# ENCYCLOPEDIA OF RESEARCH DESIGN

EDITED BY

NEIL J. SALKIND

# ENCYCLOPEDIA OF RESEARCH DESIGN

# Editorial Board


## General Editor

Neil J. Salkind
*University of Kansas*

## Associate Editors

Bruce B. Frey
*University of Kansas*

Donald M. Dougherty
*University of Texas Health Science Center at San Antonio*

## Managing Editors

Kristin Rasmussen Teasdale
*University of Kansas*

Nathalie Hill-Kapturczak
*University of Texas Health Science Center at San Antonio*

# ENCYCLOPEDIA OF RESEARCH DESIGN

VOLUME 1

EDITED BY

NEIL J. SALKIND

*University of Kansas*

Los Angeles | London | New Delhi
Singapore | Washington DC

Copyright © 2010 by SAGE Publications, Inc.

All rights reserved. No part of this book may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission in writing from the publisher.

*For information:*

SAGE Publications, Inc.
2455 Teller Road
Thousand Oaks, California 91320
E-mail: order@sagepub.com

SAGE Publications Ltd..
1 Oliver's Yard
55 City Road
London EC1Y 1SP
United Kingdom

SAGE Publications India Pvt. Ltd.
B 1/I 1 Mohan Cooperative Industrial Area
Mathura Road, New Delhi 110 044
India

SAGE Publications Asia-Pacific Pte. Ltd.
33 Pekin Street #02-01
Far East Square
Singapore 048763

Printed in the United States of America.

*Library of Congress Cataloging-in-Publication Data*

Encyclopedia of research design/edited by Neil J. Salkind.
v. cm.
Includes bibliographical references and index.
ISBN 978-1-4129-6127-1 (cloth)
1. Social sciences—Statistical methods—Encyclopedias. 2. Social sciences—Research—Methodology—Encyclopedias.
I. Salkind, Neil J.

HA29.E525 2010
001.403—dc22 2010001779

This book is printed on acid-free paper.

10 11 12 13 14 10 9 8 7 6 5 4 3 2 1

| | |
|---|---|
| *Publisher:* | Rolf A. Janke |
| *Acquisitions Editor:* | Jim Brace-Thompson |
| *Editorial Assistant:* | Michele Thompson |
| *Developmental Editor:* | Carole Maurer |
| *Reference Systems Coordinators:* | Leticia M. Gutierrez, Laura Notton |
| *Production Editor:* | Kate Schroeder |
| *Copy Editors:* | Bonnie Freeman, Liann Lech, Sheree Van Vreede |
| *Typesetter:* | C&M Digitals (P) Ltd. |
| *Proofreaders:* | Kristin Bergstad, Kevin Gleason, Sally Jaskold, Sandy Zilka |
| *Indexer:* | Virgil Diodato |
| *Cover Designer:* | Glenn Vogel |
| *Marketing Manager:* | Amberlyn McKay |

# Contents

## Volume 1

List of Entries *vii*

Reader's Guide *xiii*

About the Editors *xix*

Contributors *xxi*

Introduction *xxix*

Entries

A *1*
B *57*
C *111*
D *321*
E *399*
F *471*
G *519*

## Volume 2

List of Entries *vii*

Entries

H *561*
I *589*
J *655*
K *663*
L *681*
M *745*
N *869*
O *949*
P *985*

## Volume 3

List of Entries *vii*

Entries

Q *1149*
R *1183*
S *1295*
T *1489*
U *1583*
V *1589*
W *1611*
Y *1645*
Z *1653*

Index *1675*

# List of Entries


Abstract
Accuracy in Parameter Estimation
Action Research
Adaptive Designs in Clinical Trials
Adjusted *F* Test. *See* Greenhouse–Geisser Correction
Alternative Hypotheses
American Educational Research Association
American Psychological Association Style
American Statistical Association
Analysis of Covariance (ANCOVA)
Analysis of Variance (ANOVA)
Animal Research
Applied Research
A Priori Monte Carlo Simulation
*Aptitudes and Instructional Methods*
Aptitude-Treatment Interaction
Assent
Association, Measures of
Autocorrelation

Bar Chart
Bartlett's Test
Barycentric Discriminant Analysis
Bayes's Theorem
Behavior Analysis Design
Behrens–Fisher *t′* Statistic
Bernoulli Distribution
Beta
Between-Subjects Design. *See* Single-Subject Design; Within-Subjects Design
Bias
Biased Estimator
Bivariate Regression
Block Design
Bonferroni Procedure
Bootstrapping
Box-and-Whisker Plot
b Parameter

Canonical Correlation Analysis
Case-Only Design
Case Study
Categorical Data Analysis
Categorical Variable
Causal-Comparative Design
Cause and Effect
Ceiling Effect
Central Limit Theorem
Central Tendency, Measures of
Change Scores
Chi-Square Test
Classical Test Theory
Clinical Significance
Clinical Trial
Cluster Sampling
Coefficient Alpha
"Coefficient Alpha and the Internal Structure of Tests"
Coefficient of Concordance
Coefficient of Variation
Coefficients of Correlation, Alienation, and Determination
Cohen's *d* Statistic
Cohen's *f* Statistic
Cohen's Kappa
Cohort Design
Collinearity
Column Graph
Completely Randomized Design
Computerized Adaptive Testing
Concomitant Variable
Concurrent Validity
Confidence Intervals
Confirmatory Factor Analysis
Confounding
Congruence
Construct Validity
Content Analysis
Content Validity
Contrast Analysis
Control Group
Control Variables
Convenience Sampling
"Convergent and Discriminant Validation by the Multitrait–Multimethod Matrix"
Copula Functions
Correction for Attenuation
Correlation
Correspondence Analysis
Correspondence Principle
Covariate
C Parameter. *See* Guessing Parameter
Criterion Problem
Criterion Validity
Criterion Variable
Critical Difference
Critical Theory
Critical Thinking
Critical Value
Cronbach's Alpha. *See* Coefficient Alpha

Crossover Design
Cross-Sectional Design
Cross-Validation
Cumulative Frequency Distribution

Databases
Data Cleaning
Data Mining
Data Snooping
Debriefing
Decision Rule
Declaration of Helsinki
Degrees of Freedom
Delphi Technique
Demographics
Dependent Variable
Descriptive Discriminant Analysis
Descriptive Statistics
Dichotomous Variable
Differential Item Functioning
Directional Hypothesis
Discourse Analysis
Discriminant Analysis
Discussion Section
Dissertation
Distribution
Disturbance Terms
*Doctrine of Chances, The*
Double-Blind Procedure
Dummy Coding
Duncan's Multiple Range Test
Dunnett's Test

Ecological Validity
Effect Coding
Effect Size, Measures of
Endogenous Variables
Error
Error Rates
Estimation
Eta-Squared
Ethics in the Research Process
Ethnography
Evidence-Based Decision Making
Exclusion Criteria
Exogenous Variables
Expected Value
Experience Sampling Method
Experimental Design
Experimenter Expectancy Effect
Exploratory Data Analysis
Exploratory Factor Analysis
Ex Post Facto Study
External Validity

Face Validity
Factorial Design
Factor Loadings
False Positive
Falsifiability
Field Study
File Drawer Problem
Fisher's Least Significant Difference Test
Fixed-Effects Models
Focus Group
Follow-Up
Frequency Distribution
Frequency Table
Friedman Test
*F* Test

Gain Scores, Analysis of
Game Theory
Gauss–Markov Theorem
Generalizability Theory
General Linear Model
Graphical Display of Data
Greenhouse–Geisser Correction
Grounded Theory
Group-Sequential Designs in Clinical Trials
Growth Curve
Guessing Parameter
Guttman Scaling

Hawthorne Effect
Heisenberg Effect
Hierarchical Linear Modeling
Histogram
Holm's Sequential Bonferroni Procedure
Homogeneity of Variance
Homoscedasticity
Honestly Significant Difference (HSD) Test
Hypothesis
Inclusion Criteria
Independent Variable
Inference: Deductive and Inductive
Influence Statistics
Influential Data Points
Informed Consent
Instrumentation
Interaction
Internal Consistency Reliability
Internal Validity
Internet-Based Research Method
Interrater Reliability
Interval Scale
Intervention
Interviewing
Intraclass Correlation
Item Analysis
Item Response Theory
Item-Test Correlation

Jackknife
John Henry Effect

Kolmogorov–Smirnov Test
KR-20
Krippendorff's Alpha
Kruskal–Wallis Test
Kurtosis

L'Abbé Plot
Laboratory Experiments
Last Observation Carried Forward
Latent Growth Modeling
Latent Variable
Latin Square Design
Law of Large Numbers
Least Squares, Methods of
Levels of Measurement
Likelihood Ratio Statistic
Likert Scaling
Line Graph
LISREL
Literature Review
*Logic of Scientific Discovery, The*
Logistic Regression
Loglinear Models
Longitudinal Design

Main Effects
Mann–Whitney *U* Test
Margin of Error
Markov Chains
Matching
Matrix Algebra
Mauchly Test
MBESS
McNemar's Test
Mean
Mean Comparisons
Median
Meta-Analysis
"Meta-Analysis of Psychotherapy Outcome Studies"
Methods Section
Method Variance
Missing Data, Imputation of
Mixed- and Random-Effects Models
Mixed Methods Design
Mixed Model Design
Mode
Models
Monte Carlo Simulation
Mortality
Multilevel Modeling
Multiple Comparison Tests
Multiple Regression
Multiple Treatment Interference
Multitrait–Multimethod Matrix
Multivalued Treatment Effects
Multivariate Analysis of Variance (MANOVA)
Multivariate Normal Distribution

Narrative Research
National Council on Measurement in Education
Natural Experiments
Naturalistic Inquiry
Naturalistic Observation
Nested Factor Design
Network Analysis
Newman–Keuls Test and Tukey Test
Nominal Scale
Nomograms
Nonclassical Experimenter Effects
Nondirectional Hypotheses
Nonexperimental Design
Nonparametric Statistics
*Nonparametric Statistics for the Behavioral Sciences*
Nonprobability Sampling
Nonsignificance
Normal Distribution
Normality Assumption
Normalizing Data
Nuisance Variable
Null Hypothesis
Nuremberg Code
NVivo

Observational Research
Observations
Occam's Razor
Odds
Odds Ratio
Ogive
Omega Squared
Omnibus Tests
One-Tailed Test
"On the Theory of Scales of Measurement"
Order Effects
Ordinal Scale
Orthogonal Comparisons
Outlier
Overfitting

Pairwise Comparisons
Panel Design
Paradigm
Parallel Forms Reliability
Parameters
Parametric Statistics
Partial Correlation
Partial Eta-Squared
Partially Randomized Preference Trial Design
Participants
Path Analysis
Pearson Product-Moment Correlation Coefficient
Percentile Rank
Pie Chart
Pilot Study
Placebo
Placebo Effect
Planning Research
Poisson Distribution
Polychoric Correlation Coefficient
Polynomials
Pooled Variance
Population
Positivism
Post Hoc Analysis
Post Hoc Comparisons
Power
Power Analysis
Pragmatic Study
Precision
Predictive Validity
Predictor Variable
Pre-Experimental Design
Pretest–Posttest Design
Pretest Sensitization
Primary Data Source
Principal Components Analysis
*Probabilistic Models for Some Intelligence and Attainment Tests*
Probability, Laws of
Probability Sampling
"Probable Error of a Mean, The"
Propensity Score Analysis
Proportional Sampling
Proposal
Prospective Study
Protocol
"Psychometric Experiments"
Psychometrics
Purpose Statement
*p* Value

Q Methodology
*Q*-Statistic
Qualitative Research
Quality Effects Model
Quantitative Research
Quasi-Experimental Design
Quetelet's Index
Quota Sampling

R
$R^2$
Radial Plot
Random Assignment
Random-Effects Models
Random Error
Randomization Tests
Randomized Block Design
Random Sampling
Random Selection
Random Variable
Range
Rating
Ratio Scale
Raw Scores
Reactive Arrangements
Recruitment
Regression Artifacts
Regression Coefficient
Regression Discontinuity
Regression to the Mean
Reliability
Repeated Measures Design
Replication
Research
Research Design Principles
Research Hypothesis
Research Question
Residual Plot
Residuals
Response Bias
Response Surface Design
Restriction of Range
Results Section
Retrospective Study
Robust
Root Mean Square Error
Rosenthal Effect
Rubrics

Sample
Sample Size
Sample Size Planning
Sampling
Sampling and Retention of Underrepresented Groups
Sampling Distributions
Sampling Error
SAS
Scatterplot
Scheffé Test
Scientific Method
Secondary Data Source
Selection
Semipartial Correlation Coefficient
Sensitivity
Sensitivity Analysis
Sequence Effects
Sequential Analysis
Sequential Design
"Sequential Tests of Statistical Hypotheses"
Serial Correlation
Shrinkage
Significance, Statistical
Significance Level, Concept of
Significance Level, Interpretation and Construction
Sign Test
Simple Main Effects
Simpson's Paradox
Single-Blind Study
Single-Subject Design
Social Desirability
Software, Free
Spearman–Brown Prophecy Formula
Spearman Rank Order Correlation
Specificity
Sphericity
Split-Half Reliability
Split-Plot Factorial Design
SPSS
Standard Deviation
Standard Error of Estimate
Standard Error of Measurement
Standard Error of the Mean
Standardization
Standardized Score
Statistic
Statistica
Statistical Control
*Statistical Power Analysis for the Behavioral Sciences*
Stepwise Regression
Stratified Sampling
Structural Equation Modeling
Student's *t* Test
Sums of Squares
Survey
Survival Analysis
SYSTAT
Systematic Error
Systematic Sampling

"Technique for the Measurement of Attitudes, A"
*Teoria Statistica Delle Classi e Calcolo Delle Probabilità*
Test
Test–Retest Reliability
Theory
Theory of Attitude Measurement
Think-Aloud Methods
Thought Experiments
Threats to Validity
Thurstone Scaling
Time-Lag Study
Time-Series Study
Time Studies
Treatment(s)
Trend Analysis
Triangulation
Trimmed Mean
Triple-Blind Study
True Experimental Design
True Positive
True Score
*t* Test, Independent Samples
*t* Test, One Sample
*t* Test, Paired Samples
Tukey's Honestly Significant Difference (HSD)
Two-Tailed Test
Type I Error
Type II Error
Type III Error

Unbiased Estimator
Unit of Analysis
U-Shaped Curve

"Validity"
Validity of Measurement
Validity of Research Conclusions

Variability, Measure of
Variable
Variance
Volunteer Bias

Wave
Weber–Fechner Law
Weibull Distribution
Weights
Welch's *t* Test
Wennberg Design
White Noise
Wilcoxon Rank Sum Test
WinPepi
Winsorize
Within-Subjects Design
Yates's Correction
Yates's Notation
Yoked Control Procedure

*z* Distribution
Zelen's Randomized Consent Design
*z* Score
*z* Test

# Reader's Guide

The Reader's Guide is provided to assist readers in locating entries on related topics. It classifies entries into 28 general topical categories:

1. Descriptive Statistics
2. Distributions
3. Graphical Displays of Data
4. Hypothesis Testing
5. Important Publications
6. Inferential Statistics
7. Item Response Theory
8. Mathematical Concepts
9. Measurement Concepts
10. Organizations
11. Publishing
12. Qualitative Research
13. Reliability of Scores
14. Research Design Concepts
15. Research Designs
16. Research Ethics
17. Research Process
18. Research Validity Issues
19. Sampling
20. Scaling
21. Software Applications
22. Statistical Assumptions
23. Statistical Concepts
24. Statistical Procedures
25. Statistical Tests
26. Theories, Laws, and Principles
27. Types of Variables
28. Validity of Scores.

Entries may be listed under more than one topic.

### Descriptive Statistics

Central Tendency, Measures of
Cohen's *d* Statistic
Cohen's *f* Statistic
Correspondence Analysis
Descriptive Statistics
Effect Size, Measures of
Eta-Squared
Factor Loadings
Krippendorff's Alpha
Mean
Median
Mode
Partial Eta-Squared
Range
Standard Deviation
Statistic
Trimmed Mean
Variability, Measure of
Variance

### Distributions

Bernoulli Distribution
Copula Functions
Cumulative Frequency Distribution
Distribution
Frequency Distribution
Kurtosis
Law of Large Numbers
Normal Distribution
Normalizing Data
Poisson Distribution
Quetelet's Index
Sampling Distributions
Weibull Distribution
Winsorize
*z* Distribution

### Graphical Displays of Data

Bar Chart
Box-and-Whisker Plot
Column Graph
Frequency Table
Graphical Display of Data
Growth Curve
Histogram
L'Abbé Plot
Line Graph
Nomograms
Ogive
Pie Chart
Radial Plot
Residual Plot
Scatterplot
U-Shaped Curve

### Hypothesis Testing

Alternative Hypotheses
Beta
Critical Value
Decision Rule
Hypothesis

# Contributors


Hervé Abdi
*University of Texas at Dallas*

Mona M. Abo-Zena
*Tufts University*

J. H. Abramson
*Hebrew University*

Ashley Acheson
*University of Texas Health Science Center at San Antonio*

Alan C. Acock
*Oregon State University*

Pamela Adams
*University of Lethbridge*

Joy Adamson
*University of York*

David L. R. Affleck
*University of Montana*

Alan Agresti
*University of Florida*

James Algina
*University of Florida*

Justin P. Allen
*University of Kansas*

Terry Andres
*University of Manitoba*

Tatiyana V. Apanasovich
*Cornell University*

Jan Armstrong
*University of New Mexico*

Scott Baldwin
*Brigham Young University*

Deborah L. Bandalos
*University of Georgia*

Kimberly A. Barchard
*University of Nevada, Las Vegas*

Alyse A. Barker
*Louisiana State University*

Peter Barker
*University of Oklahoma*

J. Jackson Barnette
*Colorado School of Public Health*

Thomas Bart
*Northwestern University*

William M. Bart
*University of Minnesota*

Randy J. Bartlett
*Blue Sigma Analytics*

Philip J. Batterham
*Australian National University*

Pat Bazeley
*Australian Catholic University*

Amy S. Beavers
*University of Tennessee*

Bethany A. Bell
*University of South Carolina*

Brandon Bergman
*Nova Southeastern University*

Arjan Berkeljon
*Brigham Young University*

Panagiotis Besbeas
*Athens University of Economics and Business*

Peter Bibby
*University of Nottingham*

Tracie L. Blumentritt
*University of Wisconsin–La Crosse*

Frederick J. Boehmke
*University of Iowa*

Daniel Bolt
*University of Wisconsin*

Sara E. Bolt
*Michigan State University*

Susan J. Bondy
*University of Toronto*

Matthew J. Borneman
*Southern Illinois University*

Sarah E. Boslaugh
*Washington University*

Robert L. Boughner
*Rogers State University*

James A. Bovaird
*University of Nebraska–Lincoln*

Michelle Boyd
*Tufts University*

Clare Bradley
*Royal Holloway, University of London*

Joanna Bradley-Gilbride
*Health Psychology Research*

Wendy D. Donlin-Washington
*University of North Carolina, Wilmington*

Kristin Duncan
*San Diego State University*

Leslie A. Duram
*Southern Illinois University*

Ronald C. Eaves
*Auburn University*

Michael Eid
*Free University of Berlin*

Thomas W. Epps
*University of Virginia*

Shelley Esquivel
*University of Tennessee*

Shihe Fan
*Capital Health*

Kristen Fay
*Institute for Applied Research in Youth Development*

Christopher Finn
*University of California, Berkeley*

Ronald Fischer
*Victoria University of Wellington*

Kristin Floress
*University of Wisconsin–Stevens Point*

Eric D. Foster
*University of Iowa*

Bruce B. Frey
*University of Kansas*

Andrea E. Fritz
*Colorado State University*

Steve Fuller
*University of Warwick*

R. Michael Furr
*Wake Forest University*

John Gaber
*University of Arkansas*

Sharon L. Gaber
*University of Arkansas*

Rex Galbraith
*University College London*

Xin Gao
*York University*

Jennie K. Gill
*University of Victoria*

Steven G. Gilmour
*Queen Mary, University of London*

Jack Glaser
*University of California, Berkeley*

Perman Gochyyev
*University of Arizona*

James M. M. Good
*University of Durham*

Janna Goodwin
*Regis University*

Matthew S. Goodwin
*The Groden Center, Inc.*

Peter Goos
*Universiteit Antwerpen*

William Drew Gouvier
*Louisiana State University*

Elizabeth Grandfield
*California State University, Fullerton*

Scott Graves
*Bowling Green State University*

Leon Greene
*University of Kansas*

Elena L. Grigorenko
*Yale University*

Matthew J. Grumbein
*Kansas University*

Fei Gu
*University of Kansas*

Matthew M. Gushta
*American Institutes for Research*

Amanda Haboush
*University of Nevada, Las Vegas*

Brian D. Haig
*University of Canterbury*

Ryan Hansen
*Kansas University*

J. Michael Hardin
*University of Alabama*

Jeffrey R. Harring
*University of Maryland*

Sarah L. Hastings
*Radford University*

Curtis P. Haugtvedt
*Fisher College of Business*

Kentaro Hayashi
*University of Hawai'i*

Nancy Headlee
*University of Tennessee*

Larry V. Hedges
*Northwestern University*

Jay Hegdé
*Medical College of Georgia*

Joel M. Hektner
*North Dakota State University*

Amanda R. Hemmesch
*Brandeis University*

Chris Herrera
*Montclair State University*

David Hevey
*Trinity College Dublin*

Christiana Hilmer
*San Diego State University*

Perry R. Hinton
*Oxford Brookes University*

Bettina B. Hoeppner
*Brown University*

Scott M. Hofer
*University of Victoria*

Robert H. Horner
*University of Oregon*

David C. Howell
*University of Vermont*

Chia-Chien Hsu
*Ohio State University*

Yen-Chih Hsu
*University of Pittsburgh*

Qiaoyan Hu
*University of Illinois at Chicago*

Jason L. Huang
*Michigan State University*

Schuyler W. Huck
*University of Tennessee*

Tania B. Huedo-Medina
*University of Connecticut*

Craig R. Hullett
*University of Arizona*

David L. Hussey
*Kent State University*

Alan Hutson
*University at Buffalo*

Robert P. Igo, Jr.
*Case Western Reserve University*

Ching-Kang Ing
*Academia Sinica*

Heide Deditius Island
*Pacific University*

Lisa M. James
*University of Texas Health Science Center at San Antonio*

Samantha John
*University of Texas Health Science Center at San Antonio*

Ruthellen Josselson
*Fielding Graduate University*

Laura M. Justice
*Ohio State University*

Patricia Thatcher Kantor
*Florida State University, Florida Center for Reading Research*

George Karabatsos
*University of Illinois at Chicago*

Michael Karson
*University of Denver*

Maria Kateri
*University of Piraeus*

Kentaro Kato
*University of Minnesota*

Michael W. Kattan
*Cleveland Clinic*

Jerome P. Keating
*University of Texas at San Antonio*

Lisa A. Keller
*University of Massachusetts Amherst*

Ken Kelley
*University of Notre Dame*

John C. Kern II
*Duquesne University*

Sarah Kershaw
*Florida State University*

H. J. Keselman
*University of Manitoba*

Kristina Keyton
*Texas Tech University*

Eun Sook Kim
*Texas A&M University*

Jee-Seon Kim
*University of Wisconsin–Madison*

Seock-Ho Kim
*University of Georgia*

Seong-Hyeon Kim
*Fuller Theological Seminary*

Seoung Bum Kim
*University of Texas at Arlington*

Bruce M. King
*Clemson University*

Neal Kingston
*University of Kansas*

Roger E. Kirk
*Baylor University*

Alan J. Klockars
*University of Washington*

Takis Konstantopoulos
*Heriot-Watt University, Pennsylvania State University*

Margaret Bull Kovera
*City University of New York*

John H. Krantz
*Hanover College*

Marie Kraska
*Auburn University*

David R. Krathwohl
*Syracuse University*

Klaus Krippendorff
*University of Pennsylvania*

Jennifer Kuhn
*University of Tennessee*

Jonna M. Kulikowich
*Pennsylvania State University*

Kevin A. Kupzyk
*University of Nebraska–Lincoln*

Oi-man Kwok
*Texas A&M University*

Chiraz Labidi
*United Arab Emirates University*

Michelle Lacey
*Tulane University*

Tze Leung Lai
*Stanford University*

David Mark Lane
*Rice University*

Roger Larocca
*Oakland University*

Robert E. Larzelere
*Oklahoma State University*

Lauren A. Lee
*University of Arizona*

Marvin Lee
*Tennessee State University*

Pierre Legendre
*Université de Montréal*

Sylvie Mrug
*University of Alabama at Birmingham*

Karen D. Multon
*University of Kansas*

Daniel J. Mundfrom
*University of Northern Colorado*

Daniel L. Murphy
*University of Texas at Austin*

Mandi Wilkes Musso
*Louisiana State University*

Raymond S. Nickerson
*Tufts University*

Adelheid A. M. Nicol
*Royal Military College*

Forrest Wesron Nutter, Jr.
*Iowa State University*

Thomas G. O'Connor
*University of Rochester Medical Center*

Stephen Olejnik
*University of Georgia*

Aline Orr
*University of Texas at Austin*

Rhea L. Owens
*University of Kansas*

Serkan Ozel
*Texas A&M University*

Anita Pak
*University of Toronto*

Qing Pan
*George Washington University*

Sang Hee Park
*Indiana University–Bloomington*

Carol S. Parke
*Duquesne University*

Meagan M. Patterson
*University of Kansas*

Jamis J. Perrett
*Texas A&M University*

Indeira Persaud
*St. Vincent Community College*

Nadini Persaud
*University of the West Indies at Cave Hill*

Maria M. Pertl
*Trinity College Dublin*

John V. Petrocelli
*Wake Forest University*

Shayne B. Piasta
*Ohio State University*

Andrea M. Piccinin
*Oregon State University*

Rogério M. Pinto
*Columbia University*

Steven C. Pitts
*University of Maryland, Baltimore County*

Jason D. Pole
*Pediatric Oncology Group of Ontario*

Wayne E. Pratt
*Wake Forest University*

Katherine Presnell
*Southern Methodist University*

Jesse E. Purdy
*Southwestern University*

LeAnn Grogan Putney
*University of Nevada, Las Vegas*

Weiqiang Qian
*University of California–Riverside*

Richard Race
*Roehampton University*

Philip H. Ramsey
*Queens College of City University of New York*

Alan Reifman
*Texas Tech University*

Matthew R. Reynolds
*University of Kansas*

Dawn M. Richard
*University of Texas Health Science Center at San Antonio*

Michelle M. Riconscente
*University of Southern California*

Edward E. Rigdon
*Georgia State University*

Steven Roberts
*Australian National University*

Jon E. Roeckelein
*Mesa College*

H. Jane Rogers
*University of Connecticut*

William M. Rogers
*Grand Valley State University*

Isabella Romeo
*University of Milan–Bicocca*

Lisa H. Rosen
*University of Texas at Dallas*

Deden Rukmana
*Savannah State University*

André A. Rupp
*University of Maryland*

Ehri Ryu
*Boston College*

Darrell Sabers
*University of Arizona*

Thomas W. Sager
*University of Texas at Austin*

Neil J. Salkind
*University of Kansas*

Brian A. Sandford
*Pittsburgh State University*

Annesa Flentje Santa
*University of Montana*

Yasuyo Sawaki
*Waseda University*

David A. Sbarra
*University of Arizona*

Janina L. Scarlet
*Brooklyn College*

Stefan Schmidt
*University Medical Centre Freiburg*

Vicki L. Schmitt
*University of Alabama*

C. Melanie Schuele
*Vanderbilt University*

Stanley L. Sclove
*University of Illinois at Chicago*

Chris Segrin
*University of Arizona*

Edith Seier
*East Tennessee State University*

Jane Sell
*Texas A&M University*

Richard J. Shavelson
*Stanford University*

Yu Shen
*University of Texas M. D. Anderson Cancer Center*

Alissa R. Sherry
*University of Texas at Austin*

David J. Sheskin
*Western Connecticut State University*

Towfic Shomar
*London School of Economics*

Matthias Siemer
*University of Miami*

Carlos Nunes Silva
*University of Lisbon*

Dean Keith Simonton
*University of California, Davis*

Kishore Sinha
*Birsa Agricultural University*

Stephen G. Sireci
*University of Massachusetts*

Selcuk R. Sirin
*New York University*

Timothy Sly
*Ryerson University*

Kellie M. Smith
*John Jay College of Criminal Justice, City University of New York*

Dongjiang Song
*Association for the Advance of Medical Instrumentation*

Fujian Song
*University of East Anglia*

Roy Sorensen
*Washington University in St. Louis*

Chris Spatz
*Hendrix College*

Scott A. Spaulding
*University of Oregon*

Karen M. Staller
*University of Michigan*

Henderikus J. Stam
*University of Calgary*

Jeffrey T. Steedle
*Council for Aid to Education*

David W. Stockburger
*United States Air Force Academy*

Stephen Stockton
*University of Tennessee*

Eric R. Stone
*Wake Forest University*

David L. Streiner
*University of Toronto*

Ian Stuart-Hamilton
*University of Glamorgan*

Jeffrey Stuewig
*George Mason University*

Thuntee Sukchotrat
*University of Texas at Arlington*

Tia Sukin
*University of Massachusetts Amherst*

Minghe Sun
*University of Texas at San Antonio*

Florensia F. Surjadi
*Iowa State University*

Xinyu Tang
*University of Pittsburgh*

Hisashi Tanizaki
*Kobe University*

Tish Holub Taylor
*Private Practice (psychology)*

Kristin Rasmussen Teasdale
*Christian Psychological Services*

Felix Thoemmes
*Arizona State University*

Jay C. Thomas
*Pacific University*

Nathan A. Thompson
*Assessment Systems Corporation*

Theresa A. Thorkildsen
*University of Illinois at Chicago*

Gail Tiemann
*University of Kansas*

Rocio Titiunik
*University of California, Berkeley*

Sigmund Tobias
*University at Albany, State University of New York*

David J. Torgerson
*University of York*

Carol Toris
*College of Charleston*

Francis Tuerlinckx
*University of Leuven*

Jean M. Twenge
*San Diego State University*

Marion K. Underwood
*University of Texas at Dallas*

Gerard J. P. Van Breukelen
*Maastricht University*

Brandon K. Vaughn
*University of Texas at Austin*

Eduardo Velasco
*Morgan State University*

Wayne F. Velicer
*Cancer Prevention Research Center*

Madhu Viswanathan
*University of Illinois at Urbana-Champaign*

Hoa T. Vo
*University of Texas Health Science Center at San Antonio*

Rainer vom Hofe
*University of Cincinnati*

Richard Wagner
*Florida State University*

Abdus S. Wahed
*University of Pittsburgh*

Harald Walach
*University of Northampton*

Michael J. Walk
*University of Baltimore*

David S. Wallace
*Fayetteville State University*

John Walsh
*University of Victoria*

Hong Wang
*University of Pittsburgh*

Jun Wang
*Colorado State University*

Xuebin Wang
*Shanghai University*

Rose Marie Ward
*Miami University*

Edward A. Wasserman
*University of Iowa*

Murray Webster, Jr.
*University of North Carolina–Charlotte*

Greg William Welch
*University of Kansas*

Barbara M. Wells
*University of Kansas*

Brian J. Wells
*Cleveland Clinic*

Craig Stephen Wells
*University of Massachusetts Amherst*

Stephen G. West
*Arizona State University*

David C. Wheeler
*Emory University*

K. A. S. Wickrama
*Iowa State University*

Rand R. Wilcox
*University of Southern California*

Lynne J. Williams
*University of Toronto Scarborough*

Thomas O. Williams, Jr.
*Virginia Polytechnic Institute*

John T. Willse
*University of North Carolina at Greensboro*

Victor L. Willson
*Texas A&M University*

Joachim K. Winter
*University of Munich*

Suzanne Woods-Groves
*University of Iowa*

Jiun-Yu Wu
*Texas A&M University*

Karl L. Wuensch
*East Carolina University*

Hongwei Yang
*University of Kentucky*

Jie Yang
*University of Illinois at Chicago*

Jingyun Yang
*Massachusetts General Hospital*

Feifei Ye
*University of Pittsburgh*

Z. Ebrar Yetkiner
*Texas A&M University*

Yue Yin
*University of Illinois at Chicago*

Ke-Hai Yuan
*University of Notre Dame*

Kally Yuen
*University of Melbourne*

Elaine Zanutto
*National Analysts Worldwide*

April L. Zenisky
*University of Massachusetts Amherst*

Hantao Zhang
*University of Iowa*

Zhigang Zhang
*Memorial Sloan-Kettering Cancer Center*

Shi Zhao
*University of Illinois at Chicago*

Xiaoling Zhong
*University of Notre Dame*

Linda Reichwein Zientek
*Sam Houston State University*

Jiyun Zu
*Educational Testing Service*

# Introduction

The *Encyclopedia of Research Design* is a collection of entries written by scholars in the field of research design, the discipline of how to plan and conduct empirical research, including the use of both quantitative and qualitative methods. A simple review of the Reader's Guide shows how broad the field is, including such topics as descriptive statistics, a review of important mathematical concepts, a description and discussion of the importance of such professional organizations as the American Educational Research Association and the American Statistical Association, the role of ethics in research, important inferential procedures, and much more. Two topics are especially interesting and set this collection of volumes apart from similar works: (1) a review of important research articles that have been seminal in the field and have helped determine the direction of several ideas and (2) a review of popular tools (such as software) used to analyze results. This collection of more than 500 entries includes coverage of these topics and many more.

## Process

The first step in the creation of the *Encyclopedia of Research Design* was the identification of people with the credentials and talent to perform certain tasks. The associate editors were selected on the basis of their experience and knowledge in the field of research design, and the managing editors were selected for their experience in helping manage large projects.

Once the editor selected the associate editors and managing editors, the next step was for the group to work collectively to identify and select a thorough and complete listing of the important topics in the area of research design. This was not easy because there are hundreds, if not thousands, of topics that could be selected. We tried to select those that are the most commonly used and that readers would find most useful and important to have defined and discussed. At the same time, we had to balance this selection with the knowledge that there is never enough room to include everything. Terms were included because of a general consensus that they were essential for such a work as this.

Once the initial list of possible entries was defined in draft form, it was revised to produce the set of categories and entries that you see in the Reader's Guide at the beginning of Volume 1. We ultimately wanted topics that were sufficiently technical to enlighten the naïve but educated reader, and at the same time we wanted to avoid those topics from which only a small percentage of potential readers would benefit.

As with many other disciplines, there is a great deal of overlap in terminology within research design, as well as across related disciplines. For example, the two relatively simple entries titled Descriptive Statistics and Mean have much in common and necessarily cover some of the same content (using different words because they were written by different authors), but each entry also presents a different approach to understanding the general topic of central tendency. More advanced topics such as Analysis of Variance and Repeated Measures Design also have a significant number of conceptual ideas in common. It is impossible to avoid overlap because all disciplines contain terms and ideas that are similar, which is what gives a discipline its internal order—similar ideas and such belong together. Second, offering different language and explanations (but by no means identical words) provides a more comprehensive and varied view of important ideas. That is the strength in the diversity of the list of contributors in the

*Encyclopedia of Research Design* and why it is the perfect instrument for new learners, as well as experienced researchers, to learn about new topics or just brush up on new developments.

As we worked with the ongoing and revised drafts of entries, we recruited authors to write the various entries. Part of the process of asking scholars to participate included asking for their feedback as to what should be included in the entry and what related topics should be included. The contributors were given the draft entry list and were encouraged to suggest others ideas and directions to pursue. Many of their ideas and suggestions were useful, and often new entries were added to the list. Almost until the end of the entire process of writing entries, the entry list continued to be revised.

Once the list was finalized, we assigned each one a specific length of 1,000, 2,000, or 3,000 words. This decision was based on the importance of the topic and how many words we thought would be necessary to represent it adequately. For example, the entry titled Abstract was deemed to be relatively limited, whereas we encouraged the author of Reliability, an absolutely central topic to research design, to write at least 3,000 words. As with every other step in the development of the *Encyclopedia of Research Design*, we always allowed and encouraged authors to provide feedback about the entries they were writing and nearly always agreed to their requests.

The final step was to identify authors for each of the 513 entries. We used a variety of mechanisms, including asking advisory board members to identify scholars who were experts in a particular area; consulting professional journals, books, conference presentations, and other sources to identify authors familiar with a particular topic; and drawing on the personal contacts that the editorial board members have cultivated over many years of working in this field. If potential authors felt they could not participate, we asked them to suggest someone who might be interested in writing the entry.

Once authors were confirmed, they were given explicit directions and deadlines for completing and submitting their entry. As the entries were submitted, the editorial board of the encyclopedia read them and, if necessary, requested both format and substantive changes. Once a revised entry was resubmitted, it was once again reviewed and, when acceptable, passed on to production. Notably, most entries were acceptable on initial submission.

## How to Use the *Encyclopedia of Research Design*

The *Encyclopedia of Research Design* is a collection of entries intended for the naïve, but educated, consumer. It is a reference tool for users who may be interested in learning more about a particular research technique (such as "control group" or "reliability").

Users can search the *Encyclopedia* for specific information or browse the Reader's Guide to find topics of interest. For readers who want to pursue a topic further, each entry ends with both a list of related entries in the *Encyclopedia* and a set of further readings in the literature, often including online sources.

## Acknowledgments

As editor, I have had the pleasure of working as the lead on several Sage encyclopedias. Because of the complex nature of the topics included in the *Encyclopedia of Research Design* and the associated difficulty writing about them, this was a particularly challenging project. Many of the topics are very complex and needed extra effort on the part of the editors to identify how they might be improved. Research design is a big and complex world, and it took a special effort to parse entries down to what is contained in these pages, so a great deal of thanks goes to Dr. Bruce Frey from the University of Kansas and Dr. Donald M. Dougherty from the University of Texas Health Science Center at San Antonio for their diligence, flexibility, talent, and passion for seeing this three-volume set attain a very high standard.

Our editors at Sage, Jim Brace-Thompson, senior acquisitions editor, and Rolf Janke, vice president and publisher, SAGE Reference, do what the best editors do: provide guidance and support and leave us alone to do what we do best while they keep an eye on the entire process to be sure we do not go astray.

Kristin Teasdale and Nathalie Hill-Kapturczak acted as managing editors and with great dedication and professional skill managed to find authors, see to it that documents were submitted on time,

and track progress through the use of Sage's electronic tools. It is not an understatement that this project would not have gotten done on time or run as smoothly without their assistance.

The real behind-the-scenes heroes and heroines of this entire project are the editorial and production people at Sage who made sure that all the *i*s were dotted and the (Student) *t*s crossed. Among them is Carole Mauer, senior developmental editor, who has been the most gentle of supportive and constructive colleagues, always had the answers to countless questions, and guided us in the right directions. With Carole's grace and optimism, we were ready to do what was best for the project, even when the additional work made considerable demands. Other people we would like to sincerely thank are Michele Thompson, Leticia M. Gutierrez, Laura Notton, Kate Schroeder, Bonnie Freeman, Liann Lech, and Sheree Van Vreede, all of whom played a major role in seeing this set of volumes come to fruition. It is no exaggeration that what you see here would not have been possible without their hard work.

Of course this encyclopedia would not exist without the unselfish contributions of the many authors. They understood the task at hand was to introduce educated readers such as you to this very broad field of research design. Without exception, they performed this task admirably. While reviewing submissions, we editors would often find superb explications of difficult topics, and we became ever more pleased to be a part of this important project.

And as always, we want to dedicate this encyclopedia to our loved ones—partners, spouses, and children who are always there for us and help us see the forest through the trees, the bigger picture that makes good things great.

*Neil J. Salkind, Editor*
*University of Kansas*

*Bruce B. Frey, Associate Editor*
*University of Kansas*

*Donald M. Dougherty, Associate Editor*
*University of Texas Health Science Center at San Antonio*

## Abstract

An abstract is a summary of a research or a review article and includes critical information, including a complete reference to the work, its purpose, methods used, conclusions reached, and implications. For example, here is one such abstract from the *Journal of Black Psychology* authored by Timo Wandert from the University of Mainz, published in 2009 and titled "Black German Identities: Validating the Multidimensional Inventory of Black Identity."

All the above-mentioned elements are included in this abstract: the purpose, a brief review of important ideas to put the purpose into a context, the methods, the results, and the implications of the results.

> This study examines the reliability and validity of a German version of the Multidimensional Inventory of Black Identity (MIBI) in a sample of 170 Black Germans. The internal consistencies of all subscales are at least moderate. The factorial structure of the MIBI, as assessed by principal component analysis, corresponds to a high degree to the supposed underlying dimensional structure. Construct validity was examined by analyzing (a) the intercorrelations of the MIBI subscales and (b) the correlations of the subscales with external variables. Predictive validity was assessed by analyzing the correlations of three MIBI subscales with the level of intra-racial contact. All but one prediction concerning the correlations of the subscales could be confirmed, suggesting high validity. No statistically significant negative association was observed between the Black nationalist and assimilationist ideology subscales. This result is discussed as a consequence of the specific social context Black Germans live in and is not considered to lower the MIBI's validity. Observed differences in mean scores to earlier studies of African American racial identity are also discussed.

Abstracts serve several purposes. First, they provide a quick summary of the complete publication that is easily accessible in the print form of the article or through electronic means. Second, they become the target for search tools and often provide an initial screening when a researcher is doing a literature review. It is for this reason that article titles and abstracts contain key words that one would look for when searching for such information. Third, they become the content of reviews or collections of abstracts such as PsycINFO, published by the American Psychological Association (APA). Finally, abstracts sometimes are used as stand-ins for the actual papers when there are time or space limitations, such as at professional meetings. In this instance, abstracts are usually presented as posters in presentation sessions.

Most scholarly publications have very clear guidelines as to how abstracts are to be created, prepared, and used. For example, the APA, in the

*Publication Manual of the American Psychological Association,* provides information regarding the elements of a good abstract and suggestions for creating one. While guidelines for abstracts of scholarly publications (such as print and electronic journals) tend to differ in the specifics, the following four guidelines apply generally:

1. The abstract should be short. For example, APA limits abstracts to 250 words, and MEDLINE limits them to no more than 400 words. The abstract should be submitted as a separate page.
2. The abstract should appear as one unindented paragraph.
3. The abstract should begin with an introduction and then move to a very brief summary of the method, results, and discussion.
4. After the abstract, five related keywords should be listed. These keywords help make electronic searches efficient and successful.

With the advent of electronic means of creating and sharing abstracts, visual and graphical abstracts have become popular, especially in disciplines in which they contribute to greater understanding by the reader.

*Neil J. Salkind*

***See also*** American Psychological Association Style; Ethics in the Research Process; Literature Review

**Further Readings**

American Psychological Association. (2009). *Publication Manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Fletcher, R. H. (1988). Writing an abstract. *Journal of General Internal Medicine, 3*(6), 607–609.

Luhn, H. P. (1999). The automatic creation of literature abstracts. In I. Mani & M. T. Maybury (Eds.), *Advances in automatic text summarization* (pp. 15–21). Cambridge: MIT Press.

## Accuracy in Parameter Estimation

Accuracy in parameter estimation (AIPE) is an approach to sample size planning concerned with obtaining narrow confidence intervals. The standard AIPE approach yields the necessary sample size so that the expected width of a confidence interval will be sufficiently narrow. Because confidence interval width is a random variable based on data, the actual confidence interval will almost certainly differ from (e.g., be larger or smaller than) the expected confidence interval width. A modified AIPE approach allows sample size to be planned so that there will be some desired degree of assurance that the observed confidence interval will be sufficiently narrow. The standard AIPE approach addresses questions such as what size sample is necessary so that the expected width of the 95% confidence interval width will be no larger than $\omega$, where $\omega$ is the desired confidence interval width. However, the modified AIPE approach addresses questions such as what size sample is necessary so that there is $\gamma$ 100% assurance that the 95% confidence interval width will be no larger than $\omega$, where $\gamma$ is the desired value of the assurance parameter.

Confidence interval width is a way to operationalize the accuracy of the parameter estimate, holding everything else constant. Provided appropriate assumptions are met, a confidence interval consists of a set of plausible parameter values obtained from applying the confidence interval procedure to data, where the procedure yields intervals such that $(1-\alpha)100\%$ will correctly bracket the population parameter of interest, where $1-\alpha$ is the desired confidence interval coverage. Holding everything else constant, as the width of the confidence interval decreases, the range of plausible parameter values is narrowed, and thus more values can be excluded as implausible values for the parameter. In general, whenever a parameter value is of interest, not only should the point estimate itself be reported, but so too should the corresponding confidence interval for the parameter, as it is known that a point estimate almost certainly differs from the population value and does not give an indication of the degree of uncertainty with which the parameter has been estimated. Wide confidence intervals, which illustrate the uncertainty with which the parameter has been estimated, are generally undesirable. Because the direction, magnitude, and accuracy of an effect can be simultaneously evaluated with confidence intervals, it has been argued that planning a research study in an effort to obtain narrow confidence

intervals is an ideal way to improve research findings and increase the cumulative knowledge of a discipline.

Operationalizing accuracy as the observed confidence interval width is not new. In fact, writing in the 1930s, Jerzy Neyman used the confidence interval width as a measure of accuracy in his seminal work on the theory of confidence intervals, writing that the accuracy of estimation corresponding to a fixed value of $1-\alpha$ may be measured by the length of the confidence interval. Statistically, accuracy is defined as the square root of the mean square error, which is a function of precision and bias. When the bias is zero, accuracy and precision are equivalent concepts. The AIPE approach is so named because its goal is to improve the overall accuracy of estimates, and not just the precision or bias alone. Precision can often be improved at the expense of bias, which may or may not improve the accuracy. Thus, so as not to obtain estimates that are sufficiently precise but possibly more biased, the AIPE approach sets its goal of obtaining sufficiently accurate parameter estimates as operationalized by the width of the corresponding $(1-\alpha)100\%$ confidence interval.

Basing important decisions on the results of research studies is often the goal of the study. However, when an effect has a corresponding confidence interval that is wide, decisions based on such effect sizes need to be made with caution. It is entirely possible for a point estimate to be impressive according to some standard, but for the confidence limits to illustrate that the estimate is not very accurate. For example, a commonly used set of guidelines for the standardized mean difference in the behavioral, educational, and social sciences is that population standardized effect sizes of 0.2, 0.5, and 0.8 are regarded as small, medium, and large effects, respectively, following conventions established by Jacob Cohen beginning in the 1960s. Suppose that the population standardized mean difference is thought to be medium (i.e., 0.50), based on an existing theory and a review of the relevant literature. Further suppose that a researcher planned the sample size so that there would be a statistical power of .80 when the Type I error rate is set to .05, which yields a necessary sample size of 64 participants per group (128 total). In such a situation, supposing that the observed standardized mean difference was in fact exactly 0.50, the 95% confidence interval has a lower and upper limit of .147 and .851, respectively. Thus, the lower confidence limit is smaller than "small" and the upper confidence limit is larger than "large." Although there was enough statistical power (recall that sample size was planned so that power = .80, and indeed, the null hypothesis of no group mean difference was rejected, $p = .005$), in this case sample size was not sufficient from an accuracy perspective, as illustrated by the wide confidence interval.

Historically, confidence intervals were not often reported in applied research in the behavioral, educational, and social sciences, as well as in many other domains. Cohen once suggested researchers failed to report confidence intervals because their widths were "embarrassingly large." In an effort to plan sample size so as not to obtain confidence intervals that are embarrassingly large, and in fact to plan sample size so that confidence intervals are sufficiently narrow, the AIPE approach should be considered. The argument for planning sample size from an AIPE perspective is based on the desire to report point estimates and confidence intervals instead of or in addition to the results of null hypothesis significance tests. This paradigmatic shift has led to AIPE approaches to sample size planning becoming more useful than was previously the case, given the emphasis now placed on confidence intervals instead of a narrow focus on the results of null hypothesis significance tests.

Whereas the power analytic approach to sample size planning has as its goal the rejection of a false null hypothesis with some specified probability, the AIPE approach is not concerned with whether some specified null value can be rejected (i.e., is the null value outside the confidence interval limits?), making it fundamentally different from the power analytic approach. Not surprisingly, the AIPE and power analytic approaches can suggest very different values for sample size, depending on the particular goals (e.g., desired width or desired power) specified. The AIPE approach to sample size planning is able to simultaneously consider the direction of an effect (which is what the null hypothesis significance test provides), its magnitude (best and worst case scenarios based on the values of the confidence limits), and the accuracy with which the population parameter was estimated (via the width of the confidence interval).

The term *accuracy in parameter estimation* (and the acronym *AIPE*) was first used by Ken Kelley and Scott E. Maxwell in 2003 with an argument given for its widespread use in lieu of or in addition to the power analytic approach. However, the general idea of AIPE has appeared in the literature sporadically since at least the 1960s. James Algina, as well as Stephen Olejnik and Michael R. Jiroutek, contributed to similar approaches. The goal of the approach suggested by Algina is to have an estimate sufficiently close to its corresponding population value, and the goal suggested by Olejnik and Jiroutek is to simultaneously have a sufficient degree of power and confidence interval narrowness. Currently, the most extensive program for planning sample size from the AIPE perspective is R using the MBESS package.

*Ken Kelley*

***See also*** Confidence Intervals; Effect Size, Measures of; Power Analysis; Sample Size Planning

**Further Readings**

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). Hillsdale, NJ: Lawrence Erlbaum.

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology, 59,* 537–563.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher, 31,* 25–32.

# Action Research

Action research differs from conventional research methods in three fundamental ways. First, its primary goal is social change. Second, members of the study sample accept responsibility for helping resolve issues that are the focus of the inquiry. Third, relationships between researcher and study participants are more complex and less hierarchical. Most often, action research is viewed as a process of linking theory and practice in which scholar-practitioners explore a social situation by posing a question, collecting data, and testing a hypothesis through several cycles of action. The most common purpose of action research is to guide practitioners as they seek to uncover answers to complex problems in disciplines such as education, health sciences, sociology, or anthropology. Action research is typically underpinned by ideals of social justice and an ethical commitment to improve the quality of life in particular social settings. Accordingly, the goals of action research are as unique to each study as participants' contexts; both determine the type of data-gathering methods that will be used. Because action research can embrace natural *and* social science methods of scholarship, its use is not limited to either positivist or heuristic approaches. It is, as John Dewey pointed out, an *attitude* of inquiry rather than a single research methodology.

This entry presents a brief history of action research, describes several critical elements of action research, and offers cases for and against the use of action research.

## Historical Development

Although not officially credited with authoring the term *action research,* Dewey proposed five phases of inquiry that parallel several of the most commonly used action research processes, including curiosity, intellectualization, hypothesizing, reasoning, and testing hypotheses through action. This recursive process in scientific investigation is essential to most contemporary action research models. The work of Kurt Lewin is often considered seminal in establishing the credibility of action research. In anthropology, William Foote Whyte conducted early inquiry using an action research process similar to Lewin's. In health sciences, Reginald Revans renamed the process *action learning* while observing a process of social action among nurses and coal miners in the United Kingdom. In the area of emancipatory education, Paulo Freire is acknowledged as one of the first to undertake action research characterized by participant engagement in sociopolitical activities.

The hub of the action research movement shifted from North America to the United Kingdom in the late 1960s. Lawrence Stenhouse was instrumental in revitalizing its use among health care practitioners. John Elliott championed a form of educational action research in which the researcher-as-participant

takes increased responsibility for individual and collective changes in teaching practice and school improvement. Subsequently, the 1980s were witness to a surge of action research activity centered in Australia. Wilfred Carr and Stephen Kemmis authored *Becoming Critical,* and Kemmis and Robin McTaggart's *The Action Research Planner* informed much educational inquiry. Carl Glickman is often credited with a renewed North American interest in action research in the early 1990s. He advocated action research as a way to examine and implement principles of democratic governance; this interest coincided with an increasing North American appetite for postmodern methodologies such as personal inquiry and biographical narrative.

## Characteristics

### *Reflection*

Focused reflection is a key element of most action research models. One activity essential to reflection is referred to as *metacognition,* or thinking about thinking. Researchers ruminate on the research process even as they are performing the very tasks that have generated the problem and, during their work, derive solutions from an examination of data. Another aspect of reflection is circumspection, or learning-in-practice. Action research practitioners typically proceed through various types of reflection, including those that focus on technical proficiencies, theoretical assumptions, or moral or ethical issues. These stages are also described as learning *for* practice, learning *in* practice, and learning *from* practice. Learning for practice involves the inquiry-based activities of readiness, awareness, and training engaged in collaboratively by the researcher and participants. Learning in practice includes planning and implementing intervention strategies and gathering and making sense of relevant evidence. Learning from practice includes culminating activities and planning future research. Reflection is integral to the habits of thinking inherent in scientific explorations that trigger explicit action for change.

### *Iterancy*

Most action research is cyclical and continuous. The spiraling activities of planning, acting, observing, and reflecting recur during an action research study. Iterancy, as a unique and critical characteristic, can be attributed to Lewin's early conceptualization of action research as involving hypothesizing, planning, fact-finding (reconnaissance), execution, and analysis (see Figure 1).

These iterations comprise internal and external repetition referred to as *learning loops,* during which participants engage in successive cycles of collecting and making sense of data until agreement is reached on appropriate action. The result is some form of human activity or tangible document that is immediately applicable in participants' daily lives and instrumental in informing subsequent cycles of inquiry.

### *Collaboration*

Action research methods have evolved to include collaborative and negotiatory activities among various participants in the inquiry. Divisions between the roles of researchers and participants are frequently permeable; researchers are often defined as both full participants and external experts who engage in ongoing consultation with participants. Criteria for collaboration include evident structures for sharing power and voice; opportunities to construct common language and understanding among partners; an explicit code of ethics and principles; agreement regarding shared ownership of data; provisions for sustainable community involvement and action; and consideration of generative methods to assess the process's effectiveness.

The collaborative partnerships characteristic of action research serve several purposes. The first is to integrate into the research several tenets of evidence-based responsibility rather than documentation-based accountability. Research undertaken for purposes of accountability and institutional justification often enforces an external locus of control. Conversely, responsibility-based research is characterized by job-embedded, sustained opportunities for participants' involvement in change; an emphasis on the demonstration of professional learning; and frequent, authentic recognition of practitioner growth.

### *Role of the Researcher*

Action researchers may adopt a variety of roles to guide the extent and nature of their relationships

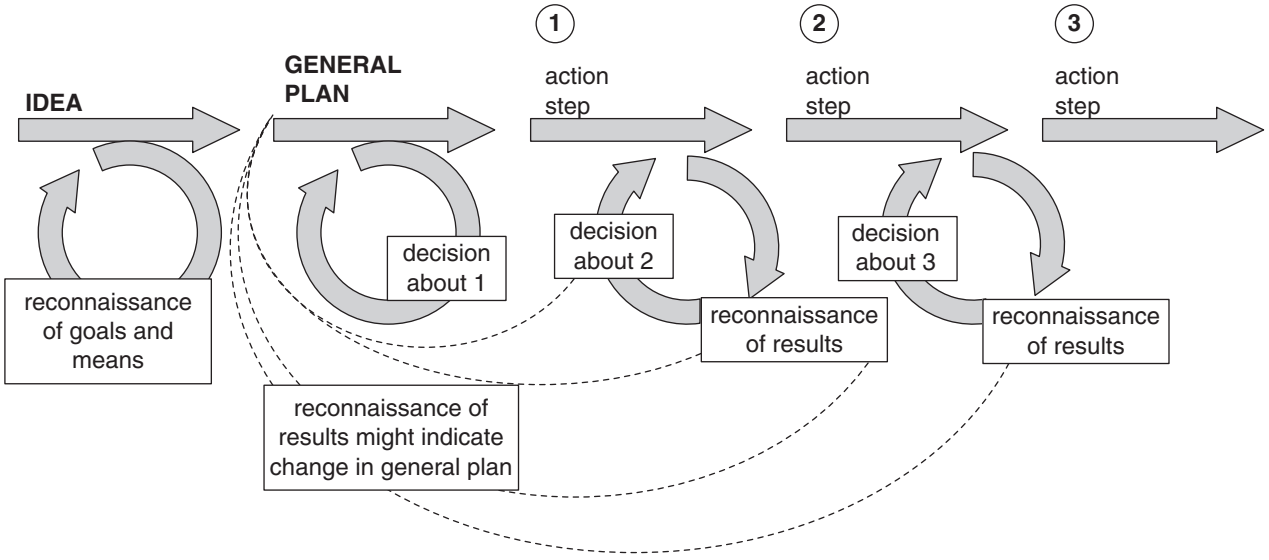



**Figure 1** Lewin's Model of Action Research

*Source:* Lewin, K. (1946). Action research and minority problems. *Journal of Social Issues, 2,* 34–46.

with participants. In a *complete participant* role, the identity of the researcher is neither concealed nor disguised. The researchers' and participants' goals are synonymous; the importance of participants' voice heightens the necessity that issues of anonymity and confidentiality are the subject of ongoing negotiation. The *participant observer* role encourages the action researcher to negotiate levels of accessibility and membership in the participant group, a process that can limit interpretation of events and perceptions. However, results derived from this type of involvement may be granted a greater degree of authenticity if participants are provided the opportunity to review and revise perceptions through a member check of observations and anecdotal data. A third possible role in action research is the *observer participant*, in which the researcher does not attempt to experience the activities and events under observation but negotiates permission to make thorough and detailed notes in a fairly detached manner. A fourth role, less common to action research, is that of the *complete observer,* in which the researcher adopts passive involvement in activities or events, and a deliberate—often physical—barrier is placed between the researcher and the participant in order to minimize contamination. These categories only hint at the complexity of roles in action research. The learning by the participants and by the researcher is rarely mutually exclusive; moreover, in practice, action researchers are most often full participants.

Intertwined purpose and the permeability of roles between the researcher and the participant are frequently elements of action research studies with agendas of emancipation and social justice. Although this process is typically one in which the external researcher is expected and required to provide some degree of expertise or advice, participants—sometimes referred to as internal researchers—are encouraged to make sense of, and apply, a wide variety of professional learning that can be translated into ethical action. Studies such as these contribute to understanding the human condition, incorporate lived experience, give public voice to experience, and expand perspectives of participant and researcher alike.

## A Case for and Against Action Research

Ontological and epistemological divisions between qualitative and quantitative approaches to research abound, particularly in debates about the credibility of action research studies. On one hand, quantitative research is criticized for drawing conclusions that are often pragmatically irrelevant; employing

methods that are overly mechanistic, impersonal, and socially insensitive; compartmentalizing, and thereby minimizing, through hypothetico-deductive schemes, the complex, multidimensional nature of human experiences; encouraging research as an isolationist and detached activity void of, and impervious to, interdependence and collaboration; and forwarding claims of objectivity that are simply not fulfilled.

On the other hand, qualitative aspects of action research are seen as quintessentially unreliable forms of inquiry because the number of uncontrolled contextual variables offers little certainty of causation. Interpretive methodologies such as narration and autobiography can yield data that are unverifiable and potentially deceptive. Certain forms of researcher involvement have been noted for their potential to unduly influence data, while some critiques contend that Hawthorne or halo effects—rather than authentic social reality—are responsible for the findings of naturalist studies.

Increased participation in action research in the latter part of the 20th century paralleled a growing demand for more pragmatic research in all fields of social science. For some humanities practitioners, traditional research was becoming irrelevant, and their social concerns and challenges were not being adequately addressed in the findings of positivist studies. They found in action research a method that allowed them to move further into other research paradigms or to commit to research that was clearly bimethodological. Increased opportunities in social policy development meant that practitioners could play a more important role in conducting the type of research that would lead to clearer understanding of social science phenomena. Further sociopolitical impetus for increased use of action research derived from the politicizing effects of the accountability movement and from an increasing solidarity in humanities professions in response to growing public scrutiny.

The emergence of action research illustrates a shift in focus from the dominance of statistical tests of hypotheses within positivist paradigms toward empirical observations, case studies, and critical interpretive accounts. Research protocols of this type are supported by several contentions, including the following:

- The complexity of social interactions makes other research approaches problematic.
- Theories derived from positivist educational research have been generally inadequate in explaining social interactions and cultural phenomena.
- Increased public examination of public institutions such as schools, hospitals, and corporate organizations requires insights of a type that other forms of research have not provided.
- Action research can provide a bridge across the perceived gap in understanding between practitioners and theorists.

### *Reliability and Validity*

The term *bias* is a historically unfriendly pejorative frequently directed at action research. As much as possible, the absence of bias constitutes conditions in which reliability and validity can increase. Most vulnerable to charges of bias are action research inquiries with a low saturation point (i.e., a small *N*), limited interrater reliability, and unclear data triangulation. Positivist studies make attempts to control external variables that may bias data; interpretivist studies contend that it is erroneous to assume that it is possible to do *any* research—particularly human science research—that is uncontaminated by personal and political sympathies and that bias can occur in the laboratory as well as in the classroom. While value-free inquiry may not exist in any research, the critical issue may not be one of credibility but, rather, one of recognizing divergent ways of answering questions associated with purpose and intent. Action research can meet determinants of reliability and validity if primary contextual variables remain consistent and if researchers are as disciplined as possible in gathering, analyzing, and interpreting the evidence of their study; in using triangulation strategies; and in the purposeful use of participation validation. Ultimately, action researchers must reflect rigorously and consistently on the places and ways that values insert themselves into studies and on how researcher tensions and contradictions can be consistently and systematically examined.

### *Generalizability*

Is any claim of replication possible in studies involving human researchers and participants?

Perhaps even more relevant to the premises and intentions that underlie action research is the question, Is this *desirable* in contributing to our understanding of the social world? Most action researchers are less concerned with the traditional goal of generalizability than with capturing the richness of unique human experience and meaning. Capturing this richness is often accomplished by reframing determinants of generalization and avoiding randomly selected examples of human experience as the basis for conclusions or extrapolations. Each instance of social interaction, if thickly described, represents a slice of the social world in the classroom, the corporate office, the medical clinic, or the community center. A certain level of generalizability of action research results may be possible in the following circumstances:

- Participants in the research recognize and confirm the accuracy of their contributions.
- Triangulation of data collection has been thoroughly attended to.
- Interrater techniques are employed prior to drawing research conclusions.
- Observation is as persistent, consistent, and longitudinal as possible.
- Dependability, as measured by an auditor, substitutes for the notion of reliability.
- Confirmability replaces the criterion of objectivity.

### *Ethical Considerations*

One profound moral issue that action researchers, like other scientists, cannot evade is the use they make of knowledge that has been generated during inquiry. For this fundamental ethical reason, the premises of any study—but particularly those of action research—must be transparent. Moreover, they must attend to a wider range of questions regarding intent and purpose than simply those of validity and reliability. These questions might include considerations such as the following:

- Why was this topic chosen?
- How and by whom was the research funded?
- To what extent does the topic dictate or align with methodology?
- Are issues of access and ethics clear?
- From what foundations are the definitions of science and truth derived?
- How are issues of representation, validity, bias, and reliability discussed?
- What is the role of the research? In what ways does this align with the purpose of the study?
- In what ways will this study contribute to knowledge and understanding?

A defensible understanding of what constitutes knowledge and of the accuracy with which it is portrayed must be able to withstand reasonable scrutiny from different perspectives. Given the complexities of human nature, complete understanding is unlikely to result from the use of a single research methodology. Ethical action researchers will make public the stance and lenses they choose for studying a particular event. With transparent intent, it is possible to honor the unique, but not inseparable, domains inhabited by social and natural, thereby accommodating appreciation for the value of multiple perspectives of the human experience.

## Making Judgment on Action Research

Action research is a relatively new addition to the repertoire of scientific methodologies, but its application and impact are expanding. Increasingly sophisticated models of action research continue to evolve as researchers strive to more effectively capture and describe the complexity and diversity of social phenomena.

Perhaps as important as categorizing action research into methodological compartments is the necessity for the researcher to bring to the study full self-awareness and disclosure of the personal and political voices that will come to bear on results and action. The action researcher must reflect on and make transparent, prior to the study, the paradoxes and problematics that will guide the inquiry and, ultimately, must do everything that is fair and reasonable to ensure that action research meets requirements of rigorous scientific study. Once research purpose and researcher intent are explicit, several alternative criteria can be used to ensure that action research is sound research. These criteria include the following types, as noted by David Scott and Robin Usher:

> *Aparadigmatic* criteria, which judge natural and social sciences by the same strategies of data

collection and which apply the same determinants of reliability and validity

*Diparadigmatic* criteria, which judge social phenomena research in a manner that is dichotomous to natural science events and which apply determinants of reliability and validity that are exclusive to social science

*Multiparadigmatic* criteria, which judge research of the social world through a wide variety of strategies, each of which employs unique postmodern determinants of social science

*Uniparadigmatic* criteria, which judge the natural and social world in ways that are redefined and reconceptualized to align more appropriately with a growing quantity and complexity of knowledge

In the final analysis, action research is favored by its proponents because it

- honors the knowledge and skills of all participants
- allows participants to be the authors of their own incremental progress
- encourages participants to learn strategies of problem solving
- promotes a culture of collaboration
- enables change to occur in context
- enables change to occur in a timely manner
- is less hierarchical and emphasizes collaboration
- accounts for rather than controls phenomena

Action research is more than reflective practice. It is a complex process that may include either qualitative or quantitative methodologies, one that has researcher and participant learning at its center. Although, in practice, action research may not often result in high levels of critical analysis, it succeeds most frequently in providing participants with intellectual experiences that are illuminative rather than prescriptive and empowering rather than coercive.

*Pamela Adams*

***See also*** Evidence-Based Decision Making; External Validity; Generalizability Theory; Mixed Methods Design; Naturalistic Inquiry

**Further Readings**

Berg, B. (2001). *Qualitative research methods for the social sciences.* Toronto, Ontario, Canada: Allyn and Bacon.

Carr, W., & Kemmis, S. (1986). *Becoming critical: Education, knowledge and action research.* Philadelphia: Farmer.

Dewey, J. (1910). *How we think*. Boston: D. C. Heath.

Freire, P. (1968). *Pedagogy of the oppressed.* New York: Herder & Herder.

Habermas, J. (1971). *Knowledge and human interests.* Boston: Beacon.

Holly, M., Arhar, J., & Kasten, W. (2005). *Action research for teachers: Traveling the yellow brick road.* Upper Saddle River, NJ: Pearson/Merrill/Prentice Hall,

Kemmis, S., & McTaggart, R. (1988). *The action research planner.* Geelong, Victoria, Australia: Deakin University.

Lewin, K. (1946). Action research and minority problems. *Journal of Social Issues, 2,* 34–46.

Revans, R. (1982). *The origins and growth of action learning.* Bromley, UK: Chartwell-Bratt.

Sagor, R. (1992). *How to conduct collaborative action research*. Alexandria, VA: Association for Supervision and Curriculum Development.

Schön, D. (1983). *The reflective practitioner.* New York: Basic Books.

## Adaptive Designs in Clinical Trials

Some designs for clinical trial research, such as drug effectiveness research, allow for modification and make use of an adaptive design. Designs such as adaptive group-sequential design, n-adjustable design, adaptive seamless phase II–III design, drop-the-loser design, adaptive randomization design, adaptive dose-escalation design, adaptive treatment-switching design, and adaptive-hypothesis design are adaptive designs.

In conducting clinical trials, investigators first formulate the research question (objectives) and then plan an adequate and well-controlled study that meets the objectives of interest. Usually, the objective is to assess or compare the effect of one or more drugs on some response. Important steps involved in the process are study design, method of analysis, selection of subjects, assignment of subjects to drugs, assessment of response, and assessment of effect in terms of hypothesis testing. All the above steps are outlined in the study protocol, and the study should follow the protocol to provide a fair and unbiased assessment of the

treatment effect. However, it is not uncommon to adjust or modify the trial, methods, or both, either at the planning stage or during the study, to provide flexibility in randomization, inclusion, or exclusion; to allow addition or exclusion of doses; to extend treatment duration; or to increase or decrease the sample size. These adjustments are mostly done for one or more of the following reasons: to increase the probability of success of the trial; to comply with budget, resource, or time constraints; or to reduce concern for safety. However, these modifications must not undermine the validity and integrity of the study. This entry defines various adaptive designs and discusses the use of adaptive designs for modifying sample size.

## Adaptive Design Variations

Adaptive design of a clinical trial is a design that allows adaptation of some aspects of the trial after its initiation without undermining the trial's validity and integrity. There are variations of adaptive designs, as described in the beginning of this entry. Here is a short description of each variation:

*Adaptive Group-Sequential Design.* Adaptive group-sequential design allows premature termination of a clinical trial on the grounds of safety, efficacy, or futility, based on interim results.

n-*Adjustable Design.* Adaptive *n*-adjustable design allows reestimation or adjustment of sample size, based on the observed data at interim.

*Adaptive Seamless Phase II–III Design.* Such a design addresses, within a single trial, objectives that are normally achieved through separate Phase IIb and Phase III trials.

*Adaptive Drop-the-Loser Design.* Adaptive drop-the-loser design allows dropping of low-performing treatment group(s).

*Adaptive Randomization Design.* Adaptive randomization design allows modification of randomization schedules.

*Adaptive Dose-Escalation Design.* An adaptive dose-escalation design is used to identify the maximum tolerated dose (MTD) of a medication. This design is usually considered optimal in later-phase clinical trials.

*Adaptive Treatment-Switching Design.* An adaptive treatment-switching design allows investigators to switch a patient's treatment from an initial assignment to an alternative treatment because of a lack of efficacy or a safety concern.

*Adaptive-Hypothesis Design.* Adaptive-hypothesis design allows change in research hypotheses based on interim analysis results.

## Sample Size

There has been considerable research on adaptive designs in which interim data at first stage are used to reestimate overall sample size. Determination of sample size for a traditional randomized clinical trial design requires specification of a clinically meaningful treatment difference, to be detected with some desired power. Such determinations can become complicated because of the need for specifying nuisance parameters such as the error variance, and the choice for a clinically meaningful treatment difference may not be straightforward. However, adjustment of sample size with proper modification of Type I error may result in an overpowered study, which wastes resources, or an underpowered study, with little chance of success.

A traditional clinical trial fixes the sample size in advance and performs the analysis after all subjects have been enrolled and evaluated. The advantages of an adaptive design over classical designs are that adaptive designs allow design assumptions (e.g., variance, treatment effect) to be modified on the basis of accumulating data and allow sample size to be modified to avoid an under- or overpowered study. However, researchers have shown that an adaptive design based on revised estimates of treatment effect is nearly always less efficient than a group sequential approach. Dramatic bias can occur when power computation is being performed because of significance of interim results. Yet medical researchers tend to prefer adaptive designs, mostly because (a) clinically meaningful effect size can change when results from other trials may suggest that smaller effects than originally postulated are meaningful; (b) it is easier to request a small budget initially, with an option to ask for supplemental funding after seeing the interim data; and (c) investigators may need to see some data before finalizing the design.

*Abdus S. Wahed*

**Further Readings**

Chow, S., Shao, J., & Wang, H. (2008). *Sample size calculations in clinical trial research*. Boca Raton, FL: Chapman & Hall.

# Adjusted *F* Test

***See*** Greenhouse–Geisser Correction

# Alternative Hypotheses

The alternative hypothesis is the hypothesis that is inferred, given a rejected *null hypothesis*. Also called the *research hypothesis*, it is best described as an explanation for why the null hypothesis was rejected. Unlike the null, the alternative hypothesis is usually of most interest to the researcher.

This entry distinguishes between two types of alternatives: the *substantive* and the *statistical*. In addition, this entry provides an example and discusses the importance of experimental controls in the inference of alternative hypotheses and the rejection of the null hypothesis.

## Substantive or Conceptual Alternative

It is important to distinguish between the substantive (or conceptual, scientific) alternative and the statistical alternative. The conceptual alternative is that which is inferred by the scientist given a rejected null. It is an explanation or theory that attempts to account for why the null was rejected. The statistical alternative, on the other hand, is simply a logical complement to the null that provides no substantive or scientific explanation as to why the null was rejected. When the null hypothesis is rejected, the statistical alternative is inferred in line with the Neyman–Pearson approach to hypothesis testing. At this point, the substantive alternative put forth by the researcher usually serves as the "reason" that the null was rejected. However, a rejected null does not by itself imply that the researcher's substantive alternative hypothesis is correct. Theoretically, there could be an infinite number of explanations for why a null is rejected.

## Example

An example can help elucidate the role of alternative hypotheses. Consider a researcher who is comparing the effects of two drugs for treating a disease. The researcher hypothesizes that one of the two drugs will be far superior in treating the disease. If the researcher rejects the null hypothesis, he or she is likely to infer that one treatment performs better than the other. In this example, the statistical alternative is a statement about the population parameters of interest (e.g., population means). When it is inferred, the conclusion is that the two means are not equal, or equivalently, that the samples were drawn from distinct populations. The researcher must then make a substantive "leap" to infer that one treatment is superior to the other. There may be many other possible explanations for the two means' not being equal; however, it is likely that the researcher will infer an alternative that is in accordance with the original purpose of the scientific study (such as wanting to show that one drug outperforms the other). It is important to remember, however, that concluding that the means are not equal (i.e., inferring the statistical alternative hypothesis) does not provide any scientific evidence at all for the chosen conceptual alternative. Particularly when it is not possible to control for all possible extraneous variables, inference of the conceptual alternative hypothesis may involve a considerable amount of guesswork, or at minimum, be heavily biased toward the interests of the researcher.

A classic example in which an incorrect alternative can be inferred is the case of the disease malaria. For many years, it was believed that the disease was caused by breathing swamp air or living around swamplands. In this case, scientists comparing samples from two populations (those who live in swamplands and those who do not) could have easily rejected the null hypothesis, which would be that the rates of malaria in the two populations were equal. They then would have inferred the statistical alternative, that the rates of malaria in the swampland population were higher. Researchers could then infer a conceptual alternative—*swamplands cause malaria*. However, without experimental control built into their study, the conceptual alternative is at best nothing more than a convenient alternative advanced by the

researchers. As further work showed, mosquitoes, which live in swampy areas, were the primary transmitters of the disease, making the swamplands alternative incorrect.

### The Importance of Experimental Control

One of the most significant challenges posed by an inference of the scientific alternative hypothesis is the infinite number of plausible explanations for the rejection of the null. There is no formal statistical procedure for arriving at the correct scientific alternative hypothesis. Researchers must rely on experimental control to help narrow the number of plausible explanations that could account for the rejection of the null hypothesis. In theory, if every conceivable extraneous variable were controlled for, then inferring the scientific alternative hypothesis would not be such a difficult task. However, since there is no way to control for every possible confounding variable (at least not in most social sciences, and even many physical sciences), the goal of good researchers must be to control for as many extraneous factors as possible. The quality and extent of experimental control is proportional to the likelihood of inferring correct scientific alternative hypotheses. Alternative hypotheses that are inferred without the prerequisite of such things as control groups built into the design of the study or experiment are at best plausible explanations as to why the null was rejected, and at worst, fashionable hypotheses that the researcher seeks to endorse without the appropriate scientific license to do so.

### Concluding Comments

Hypothesis testing is an integral part of every social science researcher's job. The statistical and conceptual alternatives are two distinct forms of the alternative hypothesis. Researchers are most often interested in the conceptual alternative hypothesis. The conceptual alternative hypothesis plays an important role; without it, no conclusions could be drawn from research (other than rejecting a null). Despite its importance, hypothesis testing in the social sciences (especially the softer social sciences) has been dominated by the desire to reject null hypotheses, whereas less attention has been focused on establishing that the correct conceptual alternative has been inferred. Surely, anyone can reject a null, but few can identify and infer a correct alternative.

*Daniel J. Denis, Annesa Flentje Santa, and Chelsea Burfeind*

***See also*** Hypothesis; Null Hypothesis

#### Further Readings

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997–1003.

Cowles, M. (2000). *Statistics in psychology: An historical perspective.* Philadelphia: Lawrence Erlbaum.

Denis, D. J. (2001). Inferring the alternative hypothesis: Risky business. *Theory & Science, 2, 1.* Retrieved December 2, 2009, from http://theoryandscience.icaap.org/content/vol002.001/03denis.html

Hays, W. L. (1994). *Statistics* (5th ed.). New York: Harcourt Brace.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference (Part 1). *Biometrika, 20A,* 175–240.

## American Educational Research Association

The American Educational Research Association (AERA) is an international professional organization based in Washington, D.C., and dedicated to promoting research in the field of education. Through conferences, publications, and awards, AERA encourages the scientific pursuit and dissemination of knowledge in the educational arena. Its membership is diverse, drawn from within the education professions, as well as from the broader social science field.

### Mission

The mission of AERA is to influence the field of education in three major ways: (1) increasing knowledge about education, (2) promoting educational research, and (3) encouraging the use of educational research results to make education better and thereby improve the common good.

## History

AERA publicizes its founding as taking place in 1916. However, its roots have been traced back to the beginnings of the educational administration research area and the school survey movement, both of which took place in the 1910s. This new spirit of cooperation between university researchers and public schools led eight individuals to found the National Association of Directors of Educational Research (NADER) as an interest group within the National Education Association's Department of Superintendence in February 1915. With the creation of its first organizational constitution in 1916, NADER committed itself to the improvement of public education through applied research. NADER's two goals were to organize educational research centers at public educational settings and to promote the use of appropriate educational measures and statistics in educational research. Full membership in this new organization was restricted to individuals who directed research bureaus, although others involved in educational research could join as associate members. NADER produced its first publication, *Educational Research Bulletin,* in 1916. Within 3 years from its founding, its membership had almost quadrupled, to 36 full members. In 1919, two of the founders of NADER started producing a new journal, the *Journal of Educational Research,* soon to be adopted by NADER as an official publication.

With the growth in educational research programs in the late 1910s to early 1920s, NADER revised its constitution in 1921 to allow full membership status for anyone involved in conducting and producing educational research, by invitation only, after approval from the executive committee. To better display this change in membership makeup, the group changed its name in 1922 to the Educational Research Association of America. This shift allowed for a large increase in membership, which grew to 329 members by 1931. Members were approximately two thirds from a university background and one third from the public schools. In 1928, the organization changed its name once more, becoming the American Educational Research Association (AERA).

After a brief uproar among AERA membership involving the ownership of the *Journal of Educational Research,* AERA decided to affiliate itself with the National Education Association in 1930, gaining Washington, D.C., offices and support for AERA's proposed new journal, the *Review of Educational Research*. AERA underwent several changes during the 1930s. Besides the creation of the new journal, AERA decided to affiliate with other professional groups that shared common interests, such as the National Committee on Research in Secondary Education and the National Council of Education. The recognition of superior research articles at an annual awards ceremony was established in 1938, although the awards ceremony was disbanded for many years starting in 1942. By 1940, AERA membership stood at 496.

Much of AERA's growth (other than membership) has come in the form of its many journals. In 1950, the *AERA Newsletter* published its first issue. Its goal was to inform the membership about current news and events in education. Its name was changed to *Educational Researcher* in 1965. In 1963, the *American Educational Research Journal* was created to give educational researchers an outlet for original research articles, as previous AERA publications focused primarily on reviews. By 1970, the *Review of Educational Research* had changed its focus, which led to the creation of another new journal, the *Review of Research in Education.* Meanwhile, membership has continued to grow, with current membership at approximately 26,000 individuals.

## Organization of Association Governance

The organization of AERA has changed since its founding to accommodate the greater membership and its diverse interests. AERA is currently governed by a council, an executive board, and standing committees. The council is responsible for policy setting and the assignment of standing committees for AERA and is formed of elected members, including the president, president-elect, immediate past president, two at-large members, division vice presidents, the chair of the special interest group (SIG) executive committee, and a graduate student representative. The council meets three times per year. In addition to serving on the council, the president appoints members of the standing committees.

The executive board is an advisory board that guides the president and executive director of

AERA. The board meets concurrently with council meetings and more frequently as needed. The board has managed elections, appointed an executive director, and selected annual meeting sites, in addition to other needed tasks. The council currently appoints 21 standing committees and charges them with conducting specific tasks in accordance with AERA policy. These committees range in focus from annual meeting policies and procedures to scholars and advocates for gender equity in education and to technology.

## Divisions

AERA has identified 12 areas of professional or academic interest to its members and has labeled these areas divisions. On joining AERA, each member selects which division to join. Members can belong to more than one division for an additional annual fee. The divisions hold business meetings and present research in their interest area at the annual meeting. The 12 divisions are as follows: Administration, Organization, and Leadership; Curriculum Studies; Learning and Instruction; Measurement and Research Methodology; Counseling and Human Development; History and Historiography; Social Context of Education; Research, Evaluation, and Assessment in Schools; Education in the Professions; Postsecondary Education; Teaching and Teacher Education; and Educational Policy and Politics.

## SIGs

SIGs are small interest groups within AERA membership. SIGs differ from divisions in that their focus tends to be on more specific topics than the broad interests represented by the divisions. Like divisions, SIGs typically hold business meetings and support presentations of research in their interest area at the annual meeting. AERA has approximately 163 SIGs. Membership in the SIGs is based on annual dues, which range from $5 to $75. SIGs range in focus from Academic Audit Research in Teacher Education to Hierarchical Linear Modeling to Rasch Measurement and to Writing and Literacies. AERA members may join as many SIGs as they wish.

## Membership

At about 26,000 members, AERA is among the largest professional organizations in the United States. Membership in AERA takes several forms but is primarily divided among voting full members and nonvoting affiliates. In order to be a voting member, either one must hold the equivalent of a master's degree or higher; be a graduate student sponsored by a voting member of one's university faculty; or have emeritus status, earned following retirement after more than 20 years of voting membership in AERA. Affiliate members are those who are interested in educational research but do not have a master's degree, undergraduate students who are sponsored by their faculty, or non–U.S. citizens who do not meet the master's degree requirement. For both voting and nonvoting memberships, students pay a reduced rate. Members of AERA gain several benefits, including a reduced cost to attend the annual meeting, free membership in one division, and free subscriptions to both *Educational Researcher* and one other AERA journal of the member's choice.

## Publications

AERA publishes six peer-reviewed journals as well as several books and a series of *Research Points,* published quarterly, which are designed to help those working on policy issues connect with current educational research findings. AERA's peer-reviewed journals include the *American Educational Research Journal, Educational Evaluation and Policy Analysis, Educational Researcher,* the *Journal of Educational and Behavioral Statistics,* the *Review of Educational Research,* and the *Review of Research in Education.*

The *American Educational Research Journal* focuses on original scientific research in the field of education. It has two subdivisions: one that examines research on teaching, learning, and human development and one for social and institutional analysis. *Educational Evaluation and Policy Analysis* publishes original research focusing on evaluation and policy analysis issues. *Educational Researcher* publishes information of general interest to a broad variety of AERA members. Interpretations and summaries of current educational research, as well as book reviews, make up the majority of its pages.

The *Journal of Educational and Behavioral Statistics* focuses on new statistical methods for use in educational research, as well as critiques of current practices. It is published jointly with the American Statistical Association. The *Review of Educational Research* publishes reviews of previously published articles by interested parties from varied backgrounds. The *Review of Research in Education* is an annual publication that solicits critical essays on a variety of topics facing the field of education. All AERA's journals are published by Sage.

## Annual Meetings

AERA's annual meetings are an opportunity to bring AERA's diverse membership together to discuss and debate the latest in educational practices and research. Approximately 16,000 attendees gather annually to listen, discuss, and learn. For the 2008 meeting, 12,024 presentation proposals were submitted, and more than 2,000 were presented. In addition to presentations, many business meetings, invited sessions, awards, and demonstrations are held. Several graduate student-oriented sessions are also held. Many sessions focusing on educational research related to the geographical location of the annual meeting are also presented. Another valuable educational opportunity is the many professional development and training courses offered during the conference. These tend to be refresher courses in statistics and research design or evaluation or workshops on new assessment tools or classroom-based activities. In addition to the scheduled sessions, exhibitors of software, books, and testing materials present their wares at the exhibit hall, and members seeking new jobs can meet prospective employers in the career center. Tours of local attractions are also available. Each year's meeting is organized around a different theme. In 2008, the annual meeting theme was Research on Schools, Neighborhoods, and Communities: Toward Civic Responsibility. The meeting takes place at the same time and place as the annual meeting of the National Council on Measurement in Education.

## Other Services and Offerings

### *Graduate Student Council*

Graduate students are supported through several programs within AERA, but the program that provides or sponsors the most offerings for graduate students is the graduate student council. This group, composed of 28 graduate students and division and staff representatives, meets at every annual meeting to plan offerings for the graduate students. Its mission is to support graduate student members to become professional researchers or practitioners though education and advocacy. The graduate student council sponsors many sessions at the annual meeting, as well as hosting a graduate student resource center at the event. It also publishes a newsletter three times per year and hosts a Listserv where graduate students can exchange information.

### *Awards*

AERA offers an extensive awards program, and award recipients are announced at the president's address during the annual meeting. AERA's divisions and SIGs also offer awards, which are presented during each group's business meeting. AERA's awards cover educational researchers at all stages of their career, from the Early Career Award to the Distinguished Contributions to Research in Education Award. Special awards are also given in other areas, including social justice issues, public service, and outstanding books.

### *Fellowships and Grants*

Several fellowships are offered through AERA, with special fellowships focusing on minority researchers, researchers interested in measurement (through a program with the Educational Testing Service), and researchers interested in large-scale studies through a partnership with the American Institutes for Research. AERA also offers several small grants for various specialties, awarded up to three times per year.

*Carol A. Carman*

*See also* American Statistical Association; National Council on Measurement in Education

### Further Readings

Hultquist, N. J. (1976). A brief history of AERA's publishing. *Educational Researcher, 5*(11), 9–13.

Mershon, S., & Schlossman, S. (2008). Education, science, and the politics of knowledge: The American Educational Research Association, 1915–1940. *American Journal of Education, 114*(3), 307–340.

**Websites**

American Educational Research Association: http://www.aera.net

# American Psychological Association Style

American Psychological Association (APA) style is a system of guidelines for writing and formatting manuscripts. APA style may be used for a number of types of manuscripts, such as theses, dissertations, reports of empirical studies, literature reviews, meta-analyses, theoretical articles, methodological articles, and case studies. APA style is described extensively in the *Publication Manual of the American Psychological Association* (APA Publication Manual). The APA Publication Manual includes recommendations on writing style, grammar, and nonbiased language, as well as guidelines for manuscript formatting, such as arrangement of tables and section headings. The first APA Publication Manual was published in 1952; the most recent edition was published in 2009. APA style is the most accepted writing and formatting style for journals and scholarly books in psychology. The use of a single style that has been approved by the leading organization in the field aids readers, researchers, and students in organizing and understanding the information presented.

## Writing Style

The APA style of writing emphasizes clear and direct prose. Ideas should be presented in an orderly and logical manner, and writing should be as concise as possible. Usual guidelines for clear writing, such as the presence of a topic sentence in each paragraph, should be followed. Previous research should be described in either the past tense (e.g., "Surles and Arthur found") or the present perfect tense (e.g., "researchers have argued"). Past tense should also be used to describe results of an empirical study conducted by the author (e.g., "self-esteem increased over time"). Present tense (e.g., "these results indicate") should be used in discussing and interpreting results and drawing conclusions.

## Nonbiased Language

APA style guidelines recommend that authors avoid language that is biased against particular groups. APA provides specific guidelines for describing age, gender, race or ethnicity, sexual orientation, and disability status. Preferred terms change over time and may also be debated within groups; authors should consult a current style manual if they are unsure of the terms that are currently preferred or considered offensive. Authors may also ask study participants which term they prefer for themselves.

General guidelines for avoiding biased language include being specific, using labels as adjectives instead of nouns (e.g., "older people" rather than "the elderly"), and avoiding labels that imply a standard of judgment (e.g., "non-White," "stroke victim").

## Formatting

The APA Publication Manual also provides a number of guidelines for formatting manuscripts. These include guidelines for use of numbers, abbreviations, quotations, and headings.

### *Tables and Figures*

Tables and figures may allow numerical information to be presented more clearly and concisely than would be possible in text. Tables and figures may also allow for greater ease in comparing numerical data (for example, the mean depression scores of experimental and control groups). Figures and tables should present information clearly and supplement, rather than restate, information provided in the text of the manuscript. Numerical data reported in a table should not be repeated in the text.

### *Headings*

Headings provide the reader with an outline of the organization of the manuscript. APA style includes five levels of heading (examples below). Most manuscripts do not require all five levels. Topics of equal importance should have the same level of heading throughout the manuscript (e.g., the Method sections of multiple experiments should have the same heading level). Having only one headed subsection within a section should be avoided.

Level 1: Centered Boldface Uppercase and Lowercase Heading

Level 2: Flush Left Boldface Uppercase and Lowercase Heading

Level 3: Indented boldface lowercase heading ending with a period.

Level 4: Indented boldface italicized lowercase heading ending with a period.

Level 5: Indented italicized lowercase heading ending with a period.

## Manuscript Sections

A typical APA style manuscript that reports on an empirical study has five sections: Abstract, Introduction, Method, Results, and Discussion.

### *Abstract*

The abstract is a concise (150–250 word) summary of the contents of a manuscript. An abstract should include a description of the topic or problem under investigation and the most important findings or conclusions. If the manuscript describes an empirical study, the abstract should also include information about participants and experimental methods. The abstract of a published article is often included in databases to allow researchers to search for relevant studies on a particular topic.

### *Introduction*

The introduction section introduces the reader to the question under investigation. The author should describe the topic or problem, discuss other research related to the topic, and state the purpose of this study. The introduction should conclude with a brief statement that states the author's hypotheses and the ways in which these hypotheses are supported by the previous research discussed in the introduction. The introduction section should demonstrate how the question at hand is grounded in theory.

### *Method*

The method section describes how the study was conducted. The method section is frequently broken up into subsections with headings such as Participants, Materials and Measures, and Procedure.

Descriptions of participants should include summaries of demographic characteristics such as participants' ages, genders, and races or ethnicities. Other demographic characteristics, such as socioeconomic status and education level, should be reported when relevant. The method by which participants were recruited (e.g., by newspaper advertisements or through a departmental subject pool) should also be included.

Materials and measures should be described such that the reader can know what would be needed to replicate the study. Measures that are commercially available or published elsewhere should be referred to by name and attributed to their authors (e.g., "Self-esteem was measured using the Perceived Competence Scale for Children (Harter, 1982)."). Measures created for the study should be described and may be reproduced in the manuscript in a table or appendix.

The procedure should describe any experimental manipulation, instructions to participants (summarized unless instructions are part of the experimental manipulation, in which case they should be presented verbatim), order in which measures and manipulations were presented, and control features (such as randomization and counterbalancing).

### *Results*

The results section presents and summarizes the data collected and discusses the statistical analyses conducted and their results. Analyses should be reported in sufficient detail to justify conclusions. All relevant analyses should be reported, even those whose results were statistically nonsignificant or

that did not support the stated hypotheses. The results section will typically include inferential statistics, such as chi-squares, $F$ tests, or $t$ tests. For these statistics, the value of the test statistic, degrees of freedom, $p$ value, and size and direction of effect should be reported, for instance, $F(1, 39) = 4.86$, $p = .04$, $\eta^2 = .12$.

The results section may include figures (such as graphs or models) and tables. Figures and tables will typically appear at the end of a manuscript. If the manuscript is being submitted for publication, notes may be included in the text to indicate where figures or tables should be placed (e.g., "Insert Table 1 here.").

### *Discussion*

The discussion section is where the findings and analyses presented in the results section are summarized and interpreted. The author should discuss the extent to which the results support the stated hypotheses. Conclusions should be drawn but should remain within the boundaries of the data obtained. Ways in which the findings of the current study relate to the theoretical perspectives presented in the introduction should also be addressed. This section should briefly acknowledge the limitations of the current study and address possible alternative explanations for the research findings. The discussion section may also address potential applications of the work or suggest future research.

## Referring to Others' Work

It is an author's job to avoid plagiarism by noting when reference is made to another's work or ideas. This obligation applies even when the author is making general statements about existing knowledge (e.g., "Self-efficacy impacts many aspects of students' lives, including achievement motivation and task persistence (Bandura, 1997)."). Citations allow a reader to be aware of the original source of ideas or data and direct the reader toward sources of additional information on a topic.

When preparing a manuscript, an author may be called on to evaluate sources and make decisions about the quality of research or veracity of claims. In general, the most authoritative sources are articles published in peer-reviewed journals (e.g., *American Psychologist, Child Development, Journal of Personality and Social Psychology*). Peer review is the process of evaluation of scientific work by other researchers with relevant areas of expertise. The methodology and conclusions of an article published in a peer-reviewed journal have been examined and evaluated by several experts in the field.

### *In-Text Citations*

Throughout the manuscript text, credit should be given to authors whose work is referenced. In-text citations allow the reader to be aware of the source of an idea or finding and locate the work in the reference list at the end of the manuscript. APA style uses an author-date citation method (e.g., Bandura, 1997). For works with one or two authors, all authors are included in each citation. For works with three to five authors, the first in-text citation lists all authors; subsequent citations list only the first author by name (e.g., Hughes, Bigler, & Levy, 2007, in first citation, Hughes et al., 2007, in subsequent citations). Works with six or more authors are always cited in the truncated et al. format.

### *Reference Lists*

References should be listed alphabetically by the last name of the first author. Citations in the reference list should include names of authors, article or chapter title, and journal or book title. References to articles in journals or other periodicals should include the article's digital object identifier if one is assigned. If a document was accessed online, the reference should include a URL (Web address) where the material can be accessed. The URL listed should be as specific as possible; for example, it should link to the article rather than to the publication's homepage. The APA Publication Manual includes guidelines for citing many different types of sources. Examples of some of the most common types of references appear below.

*Book:* Bandura, A. (1997). *Self-efficacy: The exercise of control.* New York: W. H. Freeman.

*Chapter in edited book:* Powlishta, K. K. (2004). Gender as a social category: Intergroup processes

and gender-role development. In M. Bennet & F. Sani (Eds.), *The development of the social self* (pp. 103–133). New York: Psychology Press.

*Journal article:* Hughes, J. M., Bigler, R. S., & Levy, S. R. (2007). Consequences of learning about historical racism among European American and African American children. *Child Development, 78,* 1689–1705. doi: 10.1111/j.1467–8624.2007.01096.x

*Article in periodical (magazine or newspaper):* Gladwell, M. (2006, February 6). Troublemakers: What pit bulls can teach us about profiling. *The New Yorker, 81*(46), 33–41.

*Research report:* Census Bureau. (2006). *Voting and registration in the election of November 2004.* Retrieved from U.S. Census Bureau website: http://www.census.gov/population/www/socdemo/voting.html

*Meagan M. Patterson*

***See also*** Abstract; Bias; Demographics; Discussion Section; Dissertation; Methods Section; Results Section

**Further Readings**

American Psychological Association. (2009). *Publication Manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

APA Publications and Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist, 63,* 839–851.

Carver, R. P. (1984). *Writing a publishable research report in education, psychology, and related disciplines.* Springfield, IL: Charles C Thomas.

Cuba, L. J. (2002). *A short guide to writing about social science* (4th ed.). New York: Longman.

Dunn, D. S. (2007). *A short guide to writing about psychology* (2nd ed.). New York: Longman.

Sabin, W. A. (2004). The Gregg reference manual (10th ed.). New York: McGraw-Hill.

## American Statistical Association

The American Statistical Association (ASA) is a society for scientists, statisticians, and statistics consumers representing a wide range of science and education fields. Since its inception in November 1839, the ASA has aimed to provide both statistical science professionals and the public with a standard of excellence for statistics-related projects. According to ASA publications, the society's mission is "to promote excellence in the application of statistical science across the wealth of human endeavor." Specifically, the ASA mission includes a dedication to excellence with regard to statistics in practice, research, and education; a desire to work toward bettering statistical education and the profession of statistics as a whole; a concern for recognizing and addressing the needs of ASA members; education about the proper uses of statistics; and the promotion of human welfare through the use of statistics.

Regarded as the second-oldest continuously operating professional association in the United States, the ASA has a rich history. In fact, within 2 years of its founding, the society already had a U.S. president—Martin Van Buren—among its members. Also on the list of the ASA's historical members are Florence Nightingale, Alexander Graham Bell, and Andrew Carnegie. The original founders, who united at the American Education Society in Boston to form the society, include U.S. Congressman Richard Fletcher; teacher and fundraiser William Cogswell; physician and medicine reformist John Dix Fisher; statistician, publisher, and distinguished public health author Lemuel Shattuck; and lawyer, clergyman, and poet Oliver Peabody. The founders named the new organization the American Statistical Society, a name that lasted only until the first official meeting in February 1840.

In its beginning years, the ASA developed a working relationship with the U.S. Census Bureau, offering recommendations and often lending its members as heads of the census. S. N. D. North, the 1910 president of the ASA, was also the first director of the permanent census office. The society, its membership, and its diversity in statistical activities grew rapidly after World War I as the employment of statistics in business and government gained popularity. At that time, large cities and universities began forming local chapters. By its 100th year in existence, the ASA had more members than it had ever had, and those involved with the society commemorated the

centennial with celebrations in Boston and Philadelphia. However, by the time World War II was well under way, many of the benefits the ASA experienced from the post–World War I surge were reversed. For 2 years—1942 and 1943—the society was unable to hold annual meetings. Then, after World War II, as after World War I, the ASA saw a great expansion in both its membership and its applications to burgeoning science endeavors.

Today, ASA has expanded beyond the United States and can count 18,000 individuals as members. Its members, who represent 78 geographic locations, also have diverse interests in statistics. These interests range from finding better ways to teach statistics to problem solving for homelessness and from AIDS research to space exploration, among a wide array of applications. The society comprises 24 sections, including the following: Bayesian Statistical Science, Biometrics, Biopharmaceutical Statistics, Business and Economic Statistics, Government Statistics, Health Policy Statistics, Nonparametric Statistics, Physical and Engineering Sciences, Quality and Productivity, Risk Analysis, a section for Statistical Programmers and Analysts, Statistical Learning and Data Mining, Social Statistics, Statistical Computing, Statistical Consulting, Statistical Education, Statistical Graphics, Statistics and the Environment, Statistics in Defense and National Security, Statistics in Epidemiology, Statistics in Marketing, Statistics in Sports, Survey Research Methods, and Teaching of Statistics in the Health Sciences. Detailed descriptions of each section, lists of current officers within each section, and links to each section are available on the ASA Web site.

In addition to holding meetings coordinated by more than 60 committees of the society, the ASA sponsors scholarships, fellowships, workshops, and educational programs. Its leaders and members also advocate for statistics research funding and offer a host of career services and outreach projects.

Publications from the ASA include scholarly journals, statistical magazines, books, research guides, brochures, and conference proceeding publications. Among the journals available are *American Statistician; Journal of Agricultural, Biological, and Environmental Statistics; Journal of the American Statistical Association; Journal of Business and Economic Statistics; Journal of Computational and Graphical Statistics; Journal of Educational and Behavioral Statistics; Journal of Statistical Software; Journal of Statistics Education; Statistical Analysis and Data Mining; Statistics in Biopharmaceutical Research; Statistics Surveys;* and *Technometrics*.

The official Web site of the ASA offers a more comprehensive look at the mission, history, publications, activities, and future directions of the society. Additionally, browsers can find information about upcoming meetings and events, descriptions of outreach and initiatives, the ASA bylaws and constitution, a copy of the Ethical Guidelines for Statistical Practice prepared by the Committee on Professional Ethics, and an organizational list of board members and leaders.

*Kristin Rasmussen Teasdale*

**Further Readings**

Koren, J. (1970). *The history of statistics: Their development and progress in many countries*. New York: B. Franklin. (Original work published 1918)

Mason, R. L. (1999). *ASA: The first 160 years*. Retrieved October 10, 2009, from http://www.amstat.org/about/first160years.cfm

Wilcox, W. F. (1940). Lemuel Shattuck, statistician, founder of the American Statistical Association. *Journal of the American Statistical Association, 35,* 224–235.

**Websites**

American Statistical Association: http://www.amstat.org

## Analysis of Covariance (ANCOVA)

Behavioral sciences rely heavily on experiments and quasi experiments for evaluating the effects of, for example, new therapies, instructional methods, or stimulus properties. An *experiment* includes at least two different treatments (conditions), and human participants are randomly assigned one treatment. If assignment is not based on randomization, the design is called a *quasi experiment*. The dependent variable or outcome of an experiment

or a quasi experiment, denoted by $Y$ here, is usually quantitative, such as the total score on a clinical questionnaire or the mean response time on a perceptual task. Treatments are evaluated by comparing them with respect to the mean of the outcome $Y$ using either analysis of variance (ANOVA) or analysis of covariance (ANCOVA). Multiple linear regression may also be used, and categorical outcomes require other methods, such as logistic regression. This entry explains the purposes of, and assumptions behind, ANCOVA for the classical two-group between-subjects design. ANCOVA for within-subject and split-plot designs is discussed briefly at the end.

Researchers often want to control or adjust statistically for some independent variable that is not experimentally controlled, such as gender, age, or a pretest value of $Y$. A categorical variable such as gender can be included in ANOVA as an additional factor, turning a one-way ANOVA into a two-way ANOVA. A quantitative variable such as age or a pretest recording can be included as a covariate, turning ANOVA into ANCOVA. ANCOVA is the bridge from ANOVA to multiple regression. There are two reasons for including a covariate in the analysis if it is predictive of the outcome $Y$. In randomized experiments, it reduces unexplained (within-group) outcome variance, thereby increasing the power of the treatment effect test and reducing the width of its confidence interval. In quasi experiments, it adjusts for a group difference with respect to that covariate, thereby adjusting the between-group difference on $Y$ for confounding.

## Model

The ANCOVA model for comparing two groups at posttest $Y$, using a covariate $X$, is as follows:

$$Y_{ij} = \mu + \alpha_j + \beta(X_{ij} - \overline{X}) + e_{ij}, \qquad (1)$$

where $Y_{ij}$ is the outcome for person $i$ in group $j$ (e.g., $j = 1$ for control, $j = 2$ for treated), and $X_{ij}$ is the covariate value for person $i$ in group $j$, $\mu$ is the grand mean of $Y$, $\alpha_j$ is the effect of treatment $j$, $\beta$ is the slope of the regression line for predicting $Y$ from $X$ within groups, $\overline{X}$ is the overall sample mean of covariate $X$, and $e_{ij}$ is a normally distributed residual or error term with a mean of zero and a variance $\sigma_e^2$, which is the same in both groups. By definition, $\alpha_1 + \alpha_2 = 0$, and so $\alpha_2 - \alpha_1 = 2\alpha_2$ is the expected posttest group difference adjusted for the covariate $X$. This is even better seen by rewriting Equation 1 as

$$Y_{ij} - \beta\left(X_{ij} - \overline{X}\right) = \mu + \alpha_j + e_{ij} \qquad (2)$$

showing that ANCOVA is ANOVA of $Y$ adjusted for $X$. Due to the centering of $X$, that is, the subtraction of $\overline{X}$, the adjustment is on the average zero in the total sample. So the centering affects individual outcome values and group means, but not the total or grand mean $\mu$ of $Y$.

ANCOVA can also be written as a multiple regression model:

$$Y_{ij} = \beta_0 + \beta_1 G_{ij} + \beta_2 X_{ij} + e_{ij} \qquad (3)$$

where $G_{ij}$ is a binary indicator of treatment group ($G_{i1} = 0$ for controls, $G_{i2} = 1$ for treated), and $\beta_2$ is the slope $\beta$ in Equation 3. Comparing Equation 1 with Equation 3 shows that $\beta_1 = 2\alpha_2$ and that $\beta_0 = (\mu - \alpha_2 - \beta\overline{X})$. Centering in Equation 3 both $G$ and $X$ (i.e., coding $G$ as $-1$ and $+1$, and subtracting $\overline{X}$ from $X$) will give $\beta_0 = \mu$ and $\beta_1 = \alpha_2$. Application of ANCOVA requires estimation of $\beta$ in Equation 1. Its least squares solution is $\frac{\sigma_{XY}}{\sigma_X^2}$, the within-group covariance between pre- and posttest, divided by the within-group pretest variance, which in turn are both estimated from the sample.

## Assumptions

As Equations 1 and 3 show, ANCOVA assumes that the covariate has a linear effect on the outcome and that this effect is homogeneous, the same in both groups. So there is no treatment by covariate interaction. Both the linearity and the homogeneity assumption can be tested and relaxed by adding to Equation 3 as predictors $X \times X$ and $G \times X$, respectively, but this entry concentrates on the classical model, Equation 1 or Equation 3. The assumption of homogeneity of residual variance $\sigma_e^2$ between groups can also be relaxed.

Another assumption is that $X$ is not affected by the treatment. Otherwise, $X$ must be treated as a mediator instead of as a covariate, with

consequences for the interpretation of analysis with versus without adjustment for $X$. If $X$ is measured before treatment assignment, this assumption is warranted.

A more complicated ANCOVA assumption is that $X$ is measured without error, where *error* refers to intra-individual variation across replications. This assumption will be valid for a covariate such as age but not for a questionnaire or test score, in particular not for a pretest of the outcome at hand. Measurement error in $X$ leads to *attenuation,* a decrease of its correlation with $Y$ and of its slope $\beta$ in Equation 1. This leads to a loss of power in randomized studies and to bias in nonrandomized studies.

A last ANCOVA assumption that is often mentioned, but not visible in Equation 1, is that there is no group difference on $X$. This seems to contradict one of the two purposes of ANCOVA, that is, adjustment for a group difference on the covariate. The answer is simple, however. The assumption is not required for covariates that are measured without measurement error, such as age. But if there is measurement error in $X$, then the resulting underestimation of its slope $\beta$ in Equation 1 leads to biased treatment effect estimation in case of a group difference on $X$. An exception is the case of treatment assignment based on the observed covariate value. In that case, ANCOVA is unbiased in spite of measurement error in $X$, whether groups differ on $X$ or not, and any attempt at correction for attenuation will then introduce bias. The assumption of no group difference on $X$ is addressed in more detail in a special section on the use of a pretest of the outcome $Y$ as covariate.

## Purposes

The purpose of a covariate in ANOVA depends on the design. To understand this, note that ANCOVA gives the following adjusted estimator of the group difference:

$$\hat{\Delta} = (\overline{Y}_2 - \overline{Y}_1) - \beta(\overline{X}_2 - \overline{X}_1) \qquad (4)$$

In a randomized experiment, the group difference on the covariate, $(\overline{X}_1 - \overline{X}_2)$, is zero, and so the adjusted difference $\hat{\Delta}$ is equal to the unadjusted difference $(\overline{Y}_2 - \overline{Y}_1)$, apart from sampling error. In terms of ANOVA, the mean square (*MS;* treatment) is the same with or without adjustment, again apart from sampling error. Things are different for the *MS* (error), which is the denominator of the $F$ test in ANOVA. ANCOVA estimates $\beta$ such that the *MS* (error) is minimized, thereby maximizing the power of the $F$ test. Since the standard error (*SE*) of $\hat{\Delta}$ is proportional to the square root of the *MS* (error), this *SE* is minimized, leading to more precise effect estimation by covariate adjustment.

In a nonrandomized study with groups differing on the covariate $X$, the covariate-adjusted group effect $\hat{\Delta}$ systematically differs from the unadjusted effect $(\overline{Y}_2 - \overline{Y}_1)$. It is unbiased if the ANCOVA assumptions are satisfied and treatment assignment is random conditional on the covariate, that is, random within each subgroup of persons who are homogeneous on the covariate. Although the *MS* (error) is again minimized by covariate adjustment, this does not imply that the *SE* of $\hat{\Delta}$ is reduced. This *SE* is a function not only of *MS* (error), but also of treatment–covariate correlation. In a randomized experiment, this correlation is zero apart from sampling error, and so the *SE* depends only on the *MS* (error) and sample size. In nonrandomized studies, the *SE* increases with treatment–covariate correlation and can be larger with than without adjustment. But in nonrandomized studies, the primary aim of covariate adjustment is correction for bias, not a gain of power.

The two purposes of ANCOVA are illustrated in Figures 1 and 2, showing the within-group regressions of outcome $Y$ on covariate $X$, with the ellipses summarizing the scatter of individual persons around their group line. Each group has its own regression line with the same slope $\beta$ (reflecting absence of interaction) but different intercepts. In Figure 1, of a nonrandomized study, the groups differ on the covariate. Moving the markers for both group means along their regression line to a common covariate value $\overline{X}$ gives the adjusted group difference $\hat{\Delta}$ on outcome $Y$, reflected by the vertical distance between the two lines, which is also the difference between both intercepts. In Figure 2, of a randomized study, the two groups have the same mean covariate value, and so unadjusted and adjusted group difference on $Y$ are the same. However, in both figures the adjustment has yet another effect, illustrated in Figure 2. The *MS* (error) of ANOVA without adjustment is the entire

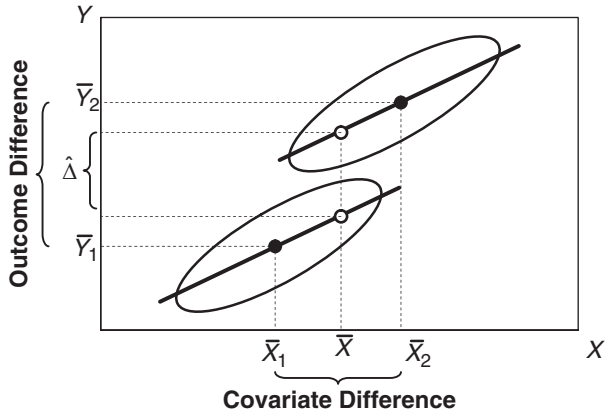



**Figure 1** Adjustment of the Outcome Difference Between Groups for a Covariate Difference in a Nonrandomized Study

*Notes:* Regression lines for treated (upper) and untreated (lower) group. Ellipses indicate scatter of individuals around their group lines. Markers on the lines indicate unadjusted (solid) and adjusted (open) group means.

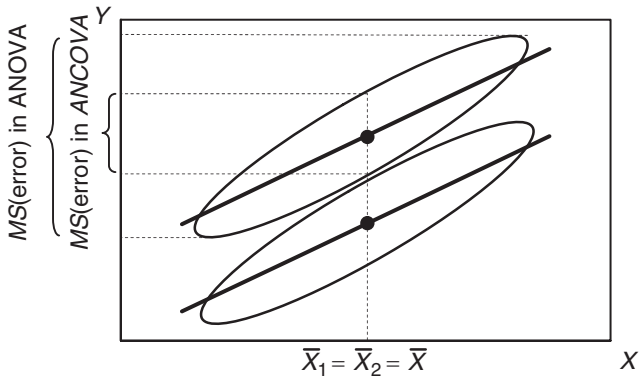



**Figure 2** Reduction of Unexplained Outcome Variance by Covariate Adjustment in a Randomized Study

*Notes:* Vertical distance between upper and lower end of an ellipse indicates unexplained (within-group) outcome variance in ANOVA. Vertical distance within an ellipse indicates unexplained outcome variance in ANCOVA.

within-group variance in vertical direction, ignoring regression lines. The *MS* (error) of ANCOVA is the variance of the vertical distances of individual dots from their group regression line. All variation in the *Y*-direction that can be predicted from the covariate; that is, all increase of *Y* along the line is included in the unadjusted *MS* (error) but excluded from the adjusted *MS* (error), which is thus smaller. In fact, it is only $\left(1-\rho_{XY}^2\right)$ as large as the unadjusted *MS* (error), where $\rho_{XY}$ is the within-group correlation between outcome and covariate.

### Using a Pretest of the Outcome as Covariate

An important special case of ANCOVA is that in which a pretest measurement of *Y* is used as covariate. The user can then choose between two methods of analysis:

1. ANCOVA with the pretest as covariate and the posttest as outcome.
2. ANOVA with the change score (posttest minus pretest) as outcome.

Two other popular methods come down to either of these two: ANCOVA of the change score is equivalent to Method 1. The Group × Time interaction test in a repeated measures ANOVA with pretest and posttest as repeated measures is equivalent to Method 2. So the choice is between Methods 1 and 2 only. Note that Method 2 is a special case of Method 1 in the sense that choosing $\beta = 1$ in ANCOVA gives ANOVA of change, as Equation 2 shows. In a randomized experiment, there is no pretest group difference, and both methods give the same unbiased treatment effect apart from sampling error, as Equation 4 shows. However, ANCOVA gives a smaller *MS* (error), leading to more test power and a smaller confidence interval than ANOVA of change, except if $\beta \approx 1$ in ANCOVA and the sample size *N* is small. In nonrandomized studies, the value for $\beta$ in Equation 4 does matter, and ANCOVA gives a different treatment effect than does ANOVA of change. The two methods may even lead to contradictory conclusions, which is known as *Lord's ANCOVA paradox*. The choice between the two methods then depends on the assignment procedure. This is best seen by writing both as a repeated measures model.

ANOVA of change is equivalent to testing the Group × Time interaction in the following model (where regression weights are denoted

by $\gamma$ to distinguish them from the $\beta s$ in earlier equations):

$$Y_{ijt} = \gamma_0 + \gamma_1 G_{ij} + \gamma_2 T_{it} + \gamma_3 G_{ij} T_{it} + e_{ijt} \qquad (5)$$

Here, $Y_{ijt}$ is the outcome value of person $i$ in group $j$ at time $t$, $G$ is the treatment group (0 = control, 1 = treated), $T$ is the time (0 = pretest, 1 = posttest), and $e_{ijt}$ is a random person effect with an unknown 2 × 2 within-group covariance matrix $\Sigma$ of pre- and posttest measures. By filling in the 0 or 1 values for $G$ and $T$, one can see that $\gamma_0$ is the pretest (population) mean of the control group, $\gamma_1$ is the pretest mean difference between the groups, $\gamma_2$ is the mean change in the control group, and $\gamma_3$ is the difference in mean change between groups. Testing the interaction effect $\gamma_3$ in Equation 5 is therefore equivalent to testing the group effect on change ($Y - X$). The only difference between repeated measures ANOVA and Equation 5 is that ANOVA uses (–1, +1) instead of (0, 1) coding for $G$ and $T$.

ANCOVA can be shown to be equivalent to testing $\gamma_3$ in Equation 5 after deleting the term $\gamma_1 G_{ij}$ by assuming $\gamma_1 = 0$, which can be done with mixed (multilevel) regression. So ANCOVA assumes that there is no group difference at pretest. This assumption is satisfied by either of two treatment assignment procedures: (1) randomization and (2) assignment based on the pretest $X$. Both designs start with one group of persons so that there can be no group effect at pretest. Groups are created after the pretest. This is why ANCOVA is the best method of analysis for both designs. In randomized experiments it has more power than ANOVA of change. With treatment assignment based on the pretest such that $\overline{X}_1 \neq \overline{X}_2$, ANCOVA is unbiased whereas ANOVA of change is then biased by ignoring regression to the mean. In contrast, if naturally occurring or preexisting groups are assigned, such as Community A getting some intervention and Community B serving as control, then ANCOVA will usually be biased whereas ANOVA of change may be unbiased. A sufficient set of conditions for ANOVA of change to be unbiased, then, is (a) that the groups are random samples from their respective populations and (b) that without treatment these populations change equally fast (or not at all). The bias in ANCOVA for this design is related to the issue of underestimation of $\beta$ in Equation 1 due to measurement error in the covariate. Correction for this underestimation gives, under certain conditions, ANOVA of change. In the end, however, the correct method of analysis for nonrandomized studies of preexisting groups is a complicated problem because of the risk of hidden confounders. Having two pretests with a suitable time interval and two control groups is then recommended to test the validity of both methods of analysis. More specifically, treating the second pretest as posttest or treating the second control group as experimental group should not yield a significant group effect because there is no treatment.

## Covariates in Other Popular Designs

This section discusses covariates in within-subject designs (e.g., *crossovers*) and between-subject designs with repeated measures (i.e., a *split-plot design*).

A within-subject design with a quantitative outcome can be analyzed with repeated measures ANOVA, which reduces to Student's paired $t$ test if there are only two treatment conditions. If a covariate such as age or a factor such as gender is added, then repeated measures ANOVA with two treatments comes down to applying ANCOVA twice: (1) to the within-subject difference $D$ of both measurements (within-subject part of the ANOVA) and (2) to the within-subject average $A$ of both measurements (between-subject part of the ANOVA). ANCOVA of $A$ tests the main effects of age and gender. ANCOVA of $D$ tests the Treatment × Gender and Treatment × Age interactions and the main effect of treatment. If gender and age are centered as in Equation 1, this main effect is $\mu$ in Equation 1, the grand mean of $D$. If gender and age are not centered, as in Equation 3, the grand mean of $D$ equals $\beta_0 + \beta_1 \overline{G} + \beta_2 \overline{X}$, where $G$ is now gender and $X$ is age. The most popular software, SPSS (an IBM company, formerly called PASW® Statistics), centers factors (here, gender) but not covariates (here, age) and tests the significance of $\beta_0$ instead of the grand mean of $D$ when reporting the $F$ test of the within-subject main effect. The optional pairwise comparison test in SPSS tests the grand mean of $D$, however.

Between-subject designs with repeated measures, for example, at posttest and follow-ups or

during and after treatment, also allow covariates. The analysis is the same as for the within-subject design extended with gender and age. But interest now is in the Treatment (between-subject) × Time (within-subject) interaction and, if there is no such interaction, in the main effect of treatment averaged across the repeated measures, rather than in the main effect of the within-subject factor time. A pretest recording can again be included as covariate or as repeated measure, depending on the treatment assignment procedure. Note, however, that as the number of repeated measures increases, the $F$ test of the Treatment × Time interaction may have low power. More powerful are the Treatment × Linear (or Quadratic) Time effect test and discriminant analysis.

Within-subject and repeated measures designs can have not only between-subject covariates such as age but also within-subject or time-dependent covariates. Examples are a baseline recording within each treatment of a crossover trial, and repeated measures of a mediator. The statistical analysis of such covariates is beyond the scope of this entry, requiring advanced methods such as mixed (multilevel) regression or structural equations modeling, although the case of only two repeated measures allows a simpler analysis by using as covariates the within-subject average and difference of the original covariate.

## Practical Recommendations for the Analysis of Studies With Covariates

Based on the preceding text, the following recommendations can be given: In randomized studies, covariates should be included to gain power, notably a pretest of the outcome. Researchers are advised to center covariates and check linearity and absence of treatment–covariate interaction as well as normality and homogeneity of variance of the residuals. In nonrandomized studies of preexisting groups, researchers should adjust for covariates that are related to the outcome to reduce bias. With two pretests or two control groups, researchers should check the validity of ANCOVA and ANOVA of change by treating the second pretest as posttest or the second control group as experimental group. No group effect should then be found. In the real posttest analysis, researchers are advised to use the average of both pretests as covariate since this average suffers less from attenuation by measurement error. In nonrandomized studies with only one pretest and one control group, researchers should apply ANCOVA and ANOVA of change and pray that they lead to the same conclusion, differing in details only.

Additionally, if there is substantial dropout related to treatment or covariates, then all data should be included in the analysis to prevent bias, using mixed (multilevel) regression instead of traditional ANOVA to prevent listwise deletion of dropouts. Further, if pretest data are used as an inclusion criterion in a nonrandomized study, then the pretest data of all excluded persons should be included in the effect analysis by mixed regression to reduce bias.

*Gerard J. P. Van Breukelen*

***See also*** Analysis of Variance (ANOVA); Covariate; Experimental Design; Gain Scores, Analysis of; Pretest–Posttest Design; Quasi-Experimental Design; Regression Artifacts; Split-Plot Factorial Design

### Further Readings

Campbell, D. T., & Kenny, D. A. (1999). *A primer on regression artifacts*. New York: Guilford Press.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Frison, L., & Pocock, S. (1997). Linearly divergent treatment effects in clinical trials with repeated measures: Efficient analysis using summary statistics. *Statistics in Medicine, 16,* 2855–2872.

Judd, C. M., Kenny, D. A., & McClelland, G. H. (2001). Estimating and testing mediation and moderation in within-subject designs. *Psychological Methods, 6,* 115–134.

Maxwell, S. E., & Delaney, H. D. (1990). *Designing experiments and analyzing data: A model comparison perspective*. Pacific Grove, CA: Brooks/Cole.

Rausch, J. R., Maxwell, S. E., & Kelley, K. (2003). Analytic methods for questions pertaining to a randomized pretest, posttest, follow-up design. *Journal of Clinical Child & Adolescent Psychology, 32*(3), 467–486.

Reichardt, C. S. (1979). The statistical analysis of data from nonequivalent group designs. In T. D. Cook & D. T. Campbell (Eds.), *Quasi-experimentation: Design*

*and analysis issues for field settings* (pp. 147–205). Boston: Houghton-Mifflin.
Rosenbaum, P. R. (1995). *Observational studies*. New York: Springer.
Senn, S. J. (2006). Change from baseline and analysis of covariance revisited. *Statistics in Medicine, 25,* 4334–4344.
Senn, S., Stevens, L., & Chaturvedi, N. (2000). Repeated measures in clinical trials: Simple strategies for analysis using summary measures. *Statistics in Medicine, 19,* 861–877.
Van Breukelen, G. J. P. (2006). ANCOVA versus change from baseline: More power in randomized studies, more bias in nonrandomized studies. *Journal of Clinical Epidemiology, 59,* 920–925.
Winkens, B., Van Breukelen, G. J. P., Schouten, H. J. A., & Berger, M. P. F. (2007). Randomized clinical trials with a pre- and a post-treatment measurement: Repeated measures versus ANCOVA models. *Contemporary Clinical Trials, 28,* 713–719.

**Table 1** Comparison of Two Treatments Based On Systolic Blood Pressure Change

| *Treatment* | |
|---|---|
| *Placebo* | *Drug A* |
| −1.3 | −4.0 |
| −1.5 | −5.7 |
| −0.5 | −3.5 |
| 0.8 | 0.4 |
| −1.1 | −1.3 |
| 3.4 | 0.8 |
| −0.8 | −10.7 |
| −3.6 | −0.3 |
| 0.3 | −0.5 |
| − 2.2 | − 3.3 |

# Analysis of Variance (ANOVA)

Usually a two-sample *t* test is applied to test for a significant difference between two population means based on the two samples. For example, consider the data in Table 1. Twenty patients with high blood pressure are randomly assigned to two groups of 10 patients. Patients in Group 1 are assigned to receive placebo, while patients in Group 2 are assigned to receive Drug A. Patients' systolic blood pressures (SBPs) are measured before and after treatment, and the differences in SBPs are recorded in Table 1. A two-sample *t* test would be an efficient method for testing the hypothesis that drug A is more effective than placebo when the differences in before and after measurements are normally distributed. However, there are usually more than two groups involved for comparison in many fields of scientific investigation. For example, extend the data in Table 1 to the data in Table 2. Here the study used 30 patients who are randomly assigned to placebo, Drug A, and Drug B. The goal here is to compare the effects of placebo and experimental drugs in reducing SBP. But a two-sample *t* test is not applicable here as we have more than two groups. Analysis of variance (ANOVA) generalizes the idea of the two-sample *t* test so that normally distributed responses can be compared across categories of one or more factors.

Since its development, ANOVA has played an indispensable role in the application of statistics in many fields, such as biology, social sciences, finance, pharmaceutics, and scientific and industrial research. Although ANOVA can be applied to various statistical models, and the simpler ones are usually named after the number of categorical variables, the concept of ANOVA is based solely on identifying the contribution of individual factors in the total variability of the data. In the above example, if the variability in SBP changes due to the drug is large compared with the chance variability, then one would think that the effect of the drug on SBP is substantial. The factors could be different individual characteristics, such as age, sex, race, occupation, social class, and treatment group, and the significant differences between the levels of these factors can be assessed by forming the ratio of the variability due to the factor itself and that due to chance only.

## History

As early as 1925, R. A. Fisher first defined the methodology of ANOVA as "separation of the variance ascribable to one group of causes from the

**Table 2** Comparison Of Three Treatments Based On Systolic Blood Pressure Change

| *Treatment* | | |
|---|---|---|
| *Placebo* | *Drug A* | *Drug B* |
| −1.3 | −4.0 | −7.6 |
| 0.5 | −5.7 | −9.2 |
| −0.5 | −3.5 | −4.0 |
| 0.4 | 0.4 | 1.8 |
| −1.1 | −1.3 | −5.3 |
| 0.6 | 0.8 | 2.6 |
| −0.8 | −10.7 | −3.8 |
| −3.6 | −0.3 | 1.2 |
| 0.3 | −0.5 | 0.4 |
| −2.2 | −3.3 | −2.6 |

variance ascribable to other groups" (p. 216). Henry Scheffé defined ANOVA as "a statistical technique for analyzing measurements depending on several kinds of effects operating simultaneously, to decide which kinds of effects are important and to estimate the effects. The measurements or observations may be in an experimental science like genetics or a nonexperimental one like astronomy" (p. 3). At first, this methodology focused more on comparing the means while treating variability as a nuisance. Nonetheless, since its introduction, ANOVA has become the most widely used statistical methodology for testing the significance of treatment effects.

Based on the number of categorical variables, ANOVA can be distinguished into one-way ANOVA and two-way ANOVA. Besides, ANOVA models can also be separated into a fixed-effects model, a random-effects model, and a mixed model based on how the factors are chosen during data collection. Each of them is described separately.

## One-Way ANOVA

One-way ANOVA is used to assess the effect of a single factor on a single response variable. When the factor is a fixed factor whose levels are the only ones of interest, one-way ANOVA is also referred to as fixed-effects one-way ANOVA. When the factor is a random factor whose levels can be considered as a sample from the population of levels, one-way ANOVA is referred to as random-effects one-way ANOVA. Fixed-effects one-way ANOVA is applied to answer the question of whether the population means are equal or not.

Given $k$ population means, the null hypothesis can be written as

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \tag{1}$$

The alternative hypothesis, $H_a$, can be written as

$H_a$ : *k population means are not all equal.*

In the random-effects one-way ANOVA model, the null hypothesis tested is that the random effect has zero variability.

Four assumptions must be met for applying ANOVA:

- A1: All samples are simple random samples drawn from each of $k$ populations representing $k$ categories of a factor.
- A2: Observations are independent of one another.
- A3: The dependent variable is normally distributed in each population.
- A4: The variance of the dependent variable is the same in each population.

Suppose, for the $j$th group, the data consist of the $n_j$ measurements $Y_{j1}, Y_{j2}, \ldots, Y_{n_j}, j = 1, 2, \ldots, k$. Then the total variation in the data can be expressed as the corrected sum of squares (*SS*) as follows: $TSS = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (Y_{ji} - \overline{y})^2$, where $\overline{y}$ is the mean of the overall sample. On the other hand, variation due to the factor is given by

$$SST = \sum_{j=1}^{k} (\overline{y}_j - \overline{y})^2, \tag{2}$$

where $\overline{y}_j$ is the mean from the $j$th group. The variation due to chance (error) is then calculated as *SSE* (error sum of squares) = *TSS* − *SST*. The component variations are usually presented in a table with corresponding degrees of freedom (*df*), mean square error, and *F* statistic. A table for one-way ANOVA is shown in Table 3.

**Table 3** General ANOVA Table for One-Way ANOVA ($k$ populations)

| *Source* | *d.f.* | *SS* | *MS* | *F* |
|---|---|---|---|---|
| *Between* | $k-1$ | *SST* | $MST = \frac{SST}{k-1}$ | $\frac{MST}{MSE}$ |
| *Within* | $n-k$ | *SSE* | $MSE = \frac{SSE}{n-k}$ | |
| *Total* | $n-1$ | *TSS* | | |

*Note:* $n$ = sample size; $k$ = number of groups; $SST$ = sum of squares treatment (factor); $MST$ = mean square treatment (factor) ; $SSE$ = sum of squares error; $TSS$ = total sum of squares.

For a given level of significance $\alpha$, the null hypothesis $H_0$ would be rejected and one could conclude that $k$ population means are not all equal if

$$F \geq F_{k-1,\, n-k,\, 1-\alpha} \tag{3}$$

where $F_{k-1,\, n-k,\, 1-\alpha}$ is the $100(1-\alpha)\%$ point of $F$ distribution with $k-1$ and $n-k$ *df*.

## Two-Way ANOVA

Two-way ANOVA is used to assess the effects of two factors and their interaction on a single response variable. There are three cases to be considered: the fixed-effects case, in which both factors are fixed; the random-effects case, in which both factors are random; and the mixed-effects case, in which one factor is fixed and the other factor is random. Two-way ANOVA is applied to answer the question of whether Factor A has a significant effect on the response adjusted for Factor B, whether Factor B has a significant effect on the response adjusted for Factor A, or whether there is an interaction effect between Factor A and Factor B.

All null hypotheses can be written as

1. $H_{01}$ : There is no Factor A effect.
2. $H_{02}$ : There is no Factor B effect.
3. $H_{03}$ : There is no interaction effect between Factor A and Factor B.

The ANOVA table for two-way ANOVA is shown in Table 4.

In the fixed case, for a given $\alpha$, the null hypothesis $H_{01}$ would be rejected, and one could conclude that there is a significant effect of Factor A if

$$F(Factor\ A) \geq F_{r-1,rc(n-1),1-\alpha}, \tag{4}$$

where $F_{r-1,\, rc(n-1),\, 1-\alpha}$ is the $100(1-\alpha)\%$ point of $F$ distribution with $r-1$ and $rc(n-1)$ *df*.

The null hypothesis $H_{02}$ would be rejected, and one could conclude that there is a significant effect of Factor B if

$$F(FactorB) \geq F_{c-1,\, rc(n-1),\, 1-\alpha} \tag{5}$$

**Table 4** General Two-Way ANOVA Table

| | | | | *F* | |
|---|---|---|---|---|---|
| *Source* | *d.f.* | *SS* | *MS* | *Fixed* | *Mixed or Random* |
| Factor A (main effect) | $r-1$ | *SSR* | $MSR = \frac{SSR}{r-1}$ | $\frac{MSR}{MSE}$ | $\frac{MSR}{MSRC}$ |
| Factor B (main effect) | $c-1$ | *SSC* | $MSC = \frac{SSC}{c-1}$ | $\frac{MSC}{MSE}$ | $\frac{MSC}{MSRC}$ |
| Factor A × Factor B (interaction) | $(r-1)(c-1)$ | *SSRC* | $MSRC = \frac{SSRC}{(r-1)(c-1)}$ | $\frac{MSRC}{MSE}$ | $\frac{MSRC}{MSE}$ |
| Error | $rc(n-1)$ | *SSE* | $MSE = \frac{SSE}{rc(n-1)}$ | | |
| Total | $rcn-1$ | *TSS* | | | |

*Note:* $r$ = number of groups for A; $c$ = number of groups for B; $SSR$ = sum of squares for Factor A; $MSR$ = mean sum of squares for Factor A; $MSRC$ = mean sum of squares for the Interaction A × B; $SSC$ = sum of squares for Factor B; $MSC$ = mean square for Factor B ; $SSRC$ = sum of squares for the Interaction $A \times B$; $SSE$ = sum of squares error; $TSS$ = total sum of squares.

where $F_{c-1,\,rc(n-1),\,1-\alpha}$ is the $100(1-\alpha)\%$ point of $F$ distribution with $c-1$ and $rc(n-1)$ *df*.

The null hypothesis $H_{03}$ would be rejected, and one could conclude that there is a significant effect of interaction between Factor A and Factor B if

$$F(Factor\,A \times Factor\,B) \geq F_{(r-1)(c-1),\,rc(n-1),\,1-\alpha}, \quad (6)$$

where $F_{(r-1)(c-1),\,rc(n-1),\,1-\alpha}$ is the $100(1-\alpha)\%$ point of $F$ distribution with $(r-1)(c-1)$ and $rc(n-1)$ *df*.

It is similar in the random case, except for different $F$ statistics and different *df* for the denominator for testing $H_{01}$ and $H_{02}$.

### Statistical Packages

SAS procedure "PROC ANOVA" performs ANOVA for balanced data from a wide variety of experimental designs. The "anova" command in STATA fits ANOVA and analysis of covariance (ANCOVA) models for balanced and unbalanced designs, including designs with missing cells; models for repeated measures ANOVA; and models for factorial, nested, or mixed designs. The "anova" function in S-PLUS produces a table with rows corresponding to each of the terms in the object, plus an additional row for the residuals. When two or more objects are used in the call, a similar table is produced showing the effects of the pairwise differences between the models, considered sequentially from the first to the last. SPSS (an IBM company, formerly called PASW® Statistics) provides a range of ANOVA options, including automated follow-up comparisons and calculations of effect size estimates.

*Abdus S. Wahed and Xinyu Tang*

***See also*** Analysis of Covariance (ANCOVA); Repeated Measures Design; *t* Test, Independent Samples

#### Further Readings

Bogartz, R. S. (1994). *An introduction to the analysis of variance.* Westport, CT: Praeger.

Fisher, R. A. (1938). *Statistical methods for research workers* (7th ed.). Edinburgh, Scotland: Oliver and Boyd.

Kleinbaum, D. G., Kupper, L. L., & Muller, K. E. (1998). *Applied regression analysis and other multivariable methods* (3rd ed.). Pacific Grove, CA: Duxbury Press.

Lindman, H. R. (1992). *Analysis of variance in experimental design.* New York: Springer-Verlag.

Scheffé, H. (1999). *Analysis of variance.* New York: Wiley-Interscience.

## Animal Research

This entry reviews the five basic research designs available to investigators who study the behavior of nonhuman animals. Use of these experimental methods is considered historically, followed by a short review of the experimental method proper. Then, for each design, the discussion focuses on manipulation of the independent variable or variables, examples of testable hypotheses, sources of error and confounding, sources of variation within the design, and statistical analyses. The entry concludes with a section on choosing the appropriate research design. In addition, this entry addresses why it is important to choose a research design prior to collecting data, why certain designs are good for testing some hypotheses but not others, and how to choose a research design. This entry focuses on nonhuman animals, but the content generalizes directly to the study of behavior, either human or nonhuman.

### The Experimental Method: Pathway to the Scientific Study of Nonhuman Animal Behavior

Through his dissertation written in 1898, Edward L. Thorndike initiated the controlled experimental analysis of nonhuman animal behavior. His use of the experimental method provided researchers interested in the evolution of intelligence, learning and memory, and mental continuity the opportunity to determine the causes of behavior. With the publication of his work in 1911 and the plea for objective methods of science by C. Lloyd Morgan, John B. Watson, and others, the use of anecdotal methods, so prevalent in that time, virtually came to an end. Thorndike's work helped establish psychology as first and foremost a science, and later a profession.

The experimental method requires at minimum two groups: the experimental group and the control group. Subjects (nonhuman animals) or participants (human animals) in the experimental group receive the treatment, and subjects or participants in the control group do not. All other variables are held constant or eliminated. When conducted correctly and carefully, the experimental method can determine cause-and-effect relationships. It is the only method that can.

## Research Designs With One Factor

### *Completely Randomized Design*

The completely randomized design is characterized by one independent variable in which subjects receive only one level of treatment. Subjects or participants are randomly drawn from a larger population, and then they are randomly assigned to one level of treatment. All other variables are held constant, counterbalanced, or eliminated. Typically, the restriction of equal numbers of subjects in each group is required. Independent variables in which subjects experience only one level are called *between-subjects variables,* and their use is widespread in the animal literature. Testable hypotheses include the following: What dosage of drug has the greatest effect on reducing seizures in rats? Which of five commercial diets for shrimp leads to the fastest growth? Does experience influence egg-laying sites in apple snails? Which of four methods of behavioral enrichment decreases abnormal behavior in captive chimpanzees the most?

The completely randomized design is chosen when *carryover effects* are of concern. Carryover effects are one form of *sequence effects* and result when the effect of one treatment level carries over into the next condition. For example, behavioral neuroscientists often lesion or ablate brain tissue to assess its role in behavioral systems including reproduction, sleep, emotion, learning, and memory. In these studies, carryover effects are almost guaranteed. Requiring all subjects to proceed through the control group first and then the experimental group is not an option. In cases in which subjects experience treatment levels in the same order, performance changes could result through practice or boredom or fatigue on the second or third or fourth time the animals experience the task. These so-called *order effects* comprise the second form of sequence effects and provide a confound wherein the experimenter does not know whether treatment effects or order effects caused the change in the dependent or response variable. *Counterbalancing* the order in which subjects receive the treatments can eliminate order effects, but in lesion studies, this is not possible. It is interesting that counterbalancing will not eliminate carryover effects. However, such effects are often, but not always, eliminated when the experimenter increases the time between conditions.

Carryover effects are not limited to a single experiment. Animal cognition experts studying sea lions, dolphins, chimpanzees, pigeons, or even gray parrots often use their subjects in multiple experiments, stretching over years. While the practice is not ideal, the cost of acquiring and maintaining the animal over its life span dictates it. Such effects can be reduced if subjects are used in experiments that differ greatly or if long periods of time have elapsed between studies. In some instances, researchers take advantage of carryover effects. Animals that are trained over long periods to perform complex tasks will often be used in an extended series of related experiments that build on this training.

Data from the completely randomized design can be statistically analyzed with parametric or nonparametric statistical tests. If the assumptions of a parametric test are met, and there are only two levels of treatment, data are analyzed with an independent $t$ test. For three or more groups, data are analyzed using the analysis of variance (ANOVA) with one between-subjects factor. Sources of variance include *treatment variance*, which includes both treatment and error variance, and *error variance* by itself. The $F$ test is Treatment + Error Variance divided by Error Variance. $F$ scores greater than 1 indicate the presence of treatment variability. Because a significant $F$ score tells the experimenter only that there is at least one significant difference, post hoc tests are required to determine where the differences lie. Depending on the experimenter's need, several post hoc tests, including a priori and a posteriori, are available. If the assumptions of a parametric test are of concern, the appropriate nonparametric test is the Mann–Whitney $U$ test for two-group designs and the Kruskal–Wallis test if three or more groups are used. Mann–Whitney $U$ tests provide post hoc analyses.

### *Repeated Measures Design*

In the repeated measures or within-subjects design, subjects experience all levels of the treatment variable, which in this case is known as the *within-subjects variable*. Subjects are randomly drawn from a larger population and exposed to all treatment conditions in orders that have been counterbalanced. All other variables have been held constant, counterbalanced, or eliminated. Examples of testable hypotheses could include the following: Is learning the reversal of a two-choice discrimination problem more difficult for a goldfish than learning the original problem? Does time of day affect the efficacy of a drug? Do bobcats choose prey on the basis of the population density of their preferred prey or the population density of their nonpreferred prey?

The repeated measures design has the major advantage of reducing the number of subjects required. For two-group designs, the number of subjects required is halved. With three or more groups, the savings are even greater. When working with animal species that are difficult to obtain, difficult to maintain, endangered, threatened, or just very expensive, this advantage is critical. A second advantage comes from having subjects experience all levels of treatment. Statistically, variability attributable to subjects can be removed from the error term. The result is that the *F* score is determined when the experimenter divides the treatment variance by residual error variance. In situations in which some subjects perform consistently at high rates of response whereas others perform consistently at low rates of response, subject variability will be high. Under these conditions, the repeated measures design is more powerful than the between-subjects design.

For example, this appears to be the case with goldfish. When goldfish are trained to strike a target for a food reward, some subjects respond at very high rates, whereas others hit the target at reduced rates. It is interesting that even in light of treatments that increase or decrease rates of response, high responders stay relatively high and low responders stay relatively low. The result is a great deal of variability that remains as error variance in a between-subjects design but that is removed from the error term in a repeated measures design. In essence, where there is considerable response bias, experimenters can use a repeated measures design to reduce the probability of making a Type II error, that is, accepting the null hypothesis as true when it is false. In the absence of response bias, the advantage of the within-subjects design is lost because the degrees of freedom are smaller relative to the between-subjects design. The important point here is that choice of research design should take into consideration the subject's traits.

The repeated measures design is subject to both carryover effects and order effects. As mentioned, counterbalancing the order that subjects receive the different levels of treatment controls for order effects. Complete counterbalancing requires that each condition both precedes and follows all other conditions an equal number of times. The equation $n(n-1)$, where $n$ equals the number of conditions, determines the number of possible orders. For example, with three treatment levels, six different orders are possible (ABC, BCA, CAB, ACB, BAC, and CBA). When the number of treatment levels increases to the point where complete counterbalancing is not possible, experimenters use *incomplete counterbalancing*. Here, only the requirement that each condition precedes all other conditions an equal number of times is followed.

Order effects can also be eliminated with a procedure commonly used in the operant conditioning literature. In the ABA design, subjects experience the control condition (A), then the experimental condition (B), then the control condition (A) again. For example, imagine a pigeon has been trained to peck a key for its favorite food. In the experiment, the pigeon first experiences a 15-minute period in which each 20th response is reinforced with 4-second access to grain (control condition). Then the pigeon receives a 15-minute period during which access time to grain differs but averages 4 seconds and the number of responses required for food also differs but averages to 20. Finally, the pigeon receives a third 15-minute period identical to the first. In this example, response rates increased dramatically during the treatment condition and then returned to the level of the first condition during the third period. When response rates return to the level observed in Condition 1, operant researchers call that *recapturing baseline*. It is often the most difficult part of operant research. Note that if the subject does not return to baseline in the third condition, the effect could

be attributable to carryover effects, order effects, or treatment effects. If the subject does return to baseline, the effect is due to the treatment and not due to sequence effects. Note also that this experiment is carried out in Las Vegas every day, except that Conditions 1 and 3 are never employed.

Data from the repeated measures design are analyzed with a dependent *t* test if two groups are used and with the one-way repeated measures ANOVA or *randomized-block design* if more than two treatment levels are used. Sources of variance include treatment variance, subject variance, and residual variance, and as mentioned, treatment variance is divided by residual variance to obtain the *F* score. Post hoc tests determine where the differences lie. The two-group nonparametric alternative is *Wilcoxon's signed ranks test*. With more than two groups, the Friedman test is chosen and Wilcoxon's signed ranks tests serve as post hoc tests.

## Research Designs With Two or More Factors

### *Completely Randomized Factorial Design*

Determining the causes of nonhuman animal behavior often requires manipulation of two or more independent variables simultaneously. Such studies are possible using factorial research designs. In the completely randomized factorial design, subjects experience only one level of each independent variable. For example, parks and wildlife managers might ask whether male or female grizzly bears are more aggressive in spring, when leaving their dens after a long period of inactivity, or in fall, when they are preparing to enter their dens. In this design, bears would be randomly chosen from a larger population and randomly assigned to one of four combinations (female fall, female spring, male fall, or male spring). The completely randomized design is used when sequence effects are of concern or if it is unlikely that the same subjects will be available for all conditions. Examples of testable hypotheses include the following: Do housing conditions of rhesus monkeys bred for medical research affect the efficacy of antianxiety drugs? Do different dosages of a newly developed pain medication affect females and males equally? Does spatial memory ability in three different species of mountain jays (corvids) change during the year? Does learning ability differ between predators and prey in marine and freshwater environments?

Besides the obvious advantage of reducing expenses by testing two or more variables at the same time, the completely randomized design can determine whether the independent variables interact to produce a third variable. Such interactions can lead to discoveries that would have been missed with a single-factor design. Consider the following example: An animal behaviorist, working for a pet food company, is asked to determine whether a new diet is good for all ages and all breeds of canines. With limited resources of time, housing, and finances, the behaviorist decides to test puppies and adults from a small breed of dogs and puppies and adults from a large breed. Twelve healthy animals for each condition (small-breed puppy, small-breed adult, large-breed puppy, and large-breed adult) are obtained from pet suppliers. All dogs are acclimated to their new surrounds for a month and then maintained on the diet for 6 months. At the end of 6 months, an index of body mass (BMI) is calculated for each dog and is subtracted from the ideal BMI determined for these breeds and ages by a group of veterinarians from the American Veterinary Medical Association. Scores of zero indicate that the diet is ideal. Negative scores reflect dogs that are gaining too much weight and becoming obese. Positive scores reflect dogs that are becoming underweight.

The results of the analysis reveal no significant effect of Age, no significant effect of Breed, but a very strong interaction. The CEO is ecstatic, believing that the absence of main effects means that the BMI scores do not differ from zero and that the new dog food is great for all dogs. However, a graph of the interaction reveals otherwise. Puppies from the small breed have very low BMI scores, indicating severely low weight. Ironically, puppies from the large breed are extremely obese, as are small-breed adult dogs. However, large-breed adults are dangerously underweight. In essence, the interaction reveals that the new diet affects small and large breeds differentially depending on age, with the outcome that the new diet would be lethal for both breeds at both age levels, but for different reasons. The lesson here is that when *main effects* are computed, the mean for the treatment level of one variable contains the

scores from all levels of the other treatment. The experimenter must examine the *interaction effect* carefully to determine whether there were main effects that were disguised by the interaction or whether there were simply no main effects.

In the end, the pet food company folded, but valuable design and statistical lessons were learned. First, in factorial designs there is always the potential for the independent variables to interact, producing a third treatment effect. Second, a significant interaction means that the independent variables affect each other differentially and that the main effects observed are confounded by the interaction. Consequently, the focus of the analysis must be on the interaction and not on the main effects. Finally, the interaction can be an expected outcome, a neutral event, or a complete surprise. For example, as discussed in the split-plot design, to demonstrate that learning has occurred, a significant interaction is required. On the surprising side, on occasion, when neuroscientists have administered two drugs at the same time, or lesioned two brain sites at the same time, the variables interacted to produce effects that relieved symptoms better than either drug alone, or the combination of lesions produced an effect never before observed.

Statistically, data from the completely randomized factorial design are analyzed with a two-way ANOVA with both factors between-subject variables. Sources of variance include treatment variability for each independent variable and each unique combination of independent variables and error variance. Error variance is estimated by adding the within-group variability for each AB cell. A single error term is used to test all treatment effects and the interaction. Post hoc tests determine where the differences lie in the main effects, and simple main effects tests are used to clarify the source of a significant interaction. Nonparametric statistical analyses for factorial designs are not routinely available on statistical packages. One is advised to check the primary literature if a nonparametric between-subjects factorial test is required.

### *The Split-Plot Design*

The split-plot or mixed design is used extensively to study animal learning and behavior. In its simplest form, the design has one between-subjects factor and one within-subjects factor. Subjects are randomly drawn from a larger population and assigned randomly to one level of the between-subjects variable. Subjects experience all levels of the within-subjects variable. The order in which subjects receive the treatment levels of the within-subjects variable is counterbalanced unless the experimenter wants to examine change over time or repeated exposures to the treatment. The split-plot design is particularly useful for studying the effects of treatment over time or exposure to treatment. Often, order of treatment cannot be counterbalanced in the within-subject factor. In these cases, order effects are expected. Examples of testable hypotheses include the following: Can crabs learn to associate the presence of polluted sand with illness? Can cephalopods solve a multiple T-maze faster than salmonids? Does tolerance to pain medication differ between males and females? Does maturation of the hippocampus in the brains of rats affect onset of the paradoxical effects of reward?

In experiments in which learning is the focus, order effects are confounded with Trials, the within-subjects treatment, and an interaction is expected. For example, in a typical classical conditioning experiment, subjects are randomly assigned to the paired group or the unpaired group. Subjects in the paired group receive paired presentations of a stimulus (light, tone, etc.) and a second stimulus that causes a response (meat powder, light foot shock). Subjects in the unpaired group receive presentations of both stimuli, but they are never paired. Group (paired/unpaired) serves as the between-subjects factor, and Trials (1–60) serves as the within-subjects factor. Evidence of learning is obtained when the number of correct responses increases as a function of trials for the paired group, but not for the unpaired group. Thus, the interaction, and not the main effects, is critical in experiments on learning.

Statistical analysis of the split-plot or mixed design is accomplished with the mixed design ANOVA, with one or more between factors and one or more within factors. With one between and one within factor, there are five sources of variance. These include treatment variance for both main effects and the interaction and two sources of error variance. The between factor is tested with error variability attributable to that factor. The within factor and the interaction are tested with residual error variance. Significant main effects are

further analyzed by post hoc tests, and the sources of the interaction are determined with simple main effects tests. Nonparametric tests are available in the primary literature.

### *Randomized Block Factorial Design*

In this design, subjects experience all levels of two or more independent variables. Such experiments can be difficult and time-consuming to conduct, to analyze, and to interpret. Sequence effects are much more likely and difficult to control. Carryover effects can result from the interaction as well as the main effects of treatment and consequently can be more difficult to detect or eliminate. Order effects can be controlled through counterbalancing, but the number of possible orders quickly escalates, as exemplified by the equation $m(n)(n-1)$, where $m$ equals the number of treatment levels in the independent variable $m$, and $n$ equals the number of treatment levels in $n$.

Because subjects experience all levels of all treatments, subject variability can be subtracted from the error term. This advantage, coupled with the need for fewer subjects and the ability to test for interactions, makes this design of value in learning experiments with exotic animals. The design also finds application in the behavioral neurosciences when the possibility of interactions between drugs presented simultaneously or sequentially needs to be assessed. The design can also assess the effects of maturation, imprinting, learning, or practice on important behavioral systems, including foraging, migration, navigation, habitat selection, choosing a mate, parental care, and so on. The following represent hypotheses that are testable with this design: Do nest site selection and nest building change depending on the success of last year's nest? Do successive operations to relieve herniated discs lead to more damage when coupled with physical therapy? Are food preferences in rhesus monkeys related to nutritional value, taste, or social learning from peers? Does predation pressure influence a prey's choice of diet more in times when food is scarce, or when food is abundant?

Data analysis is accomplished with a repeated measures factorial ANOVA. Seven sources of variance are computed: subject variance, treatment variance for both main effects and the interaction, and three error terms. Separate error terms are used to test each treatment effect and the interaction. Post hoc tests and simple main effects tests determine where the differences lie in the main effects and the interaction, respectively. Nonparametric analyses for this design can be found in the primary literature.

## Choosing a Research Design

Choosing a research design that is cost effective, feasible, and persuasive and that minimizes the probability of making Type 1 errors (experimenter rejects the null hypothesis when in fact it is true) and Type II errors (experimenter accepts the null hypothesis when in fact it is false) requires information about the availability, accessibility, maintenance, care, and cost of subjects. It also requires knowledge of the subjects themselves, including their stable traits, their developmental and evolutionary histories, their adaptability to laboratory life, their tolerance of treatment, and their ability to be trained. Knowledge of potential carryover effects associated with the treatment and whether these effects are short lived or long lasting is also important. Finally, the researcher needs to know how treatment variance and error variance are partitioned to take advantage of the traits of some animals or to increase the feasibility of conducting the experiment. In addition, the researcher needs to know how interaction effects can provide evidence of change over time or lead to new discoveries.

But there is more. Beginning-level researchers often make two critical mistakes when establishing a program of research. First, in their haste and enthusiasm, they rush out and collect data and then come back to the office not knowing how to statistically analyze their data. When these same people take their data to a statistician, they soon learn a critical lesson: *Never* conduct an experiment and then attempt to fit the data to a particular design. Choose the research design first, and then collect the data according to the rules of the design. Second, beginning-level researchers tend to think that the more complex the design, the more compelling the research. Complexity does not correlate positively with impact. Investigators should opt for the simplest design that can answer the question. The easier it is to interpret a results

section, the more likely it is that reviewers will understand and accept the findings and the conclusions. Simple designs are easier to conduct, analyze, interpret, and communicate to peers.

*Jesse E. Purdy*

***See also*** Analysis of Variance (ANOVA); Confounding; Factorial Design; Nonparametric Statistics for the Behavioral Sciences; Parametric Statistics; Post Hoc Comparisons; Single-Subject Design

**Further Readings**

Conover, W. J. (1999). *Practical nonparametric statistics* (3rd ed.). New York: Wiley.
Glover, T. J., & Mitchell, K. J. (2002). *An introduction to biostatistics*. New York: McGraw-Hill.
Howell, D. C. (2007). *Statistical methods for psychology* (6th ed.). Monterey, CA: Thomson/Wadsworth.
Kirk, R. E. (1995). *Experimental design: Procedures for behavioral sciences* (3rd ed.). Monterey, CA: Wadsworth.
Lehman, A., O'Rourke, N., Hatcher, L., & Stepanski, E. J. (2005). *JMP for basic univariate and multivariate statistics: A step-by-step guide*. Cary, NC: SAS Institute.
Wasserman, L. (2005). *All of nonparametric statistics*. New York: Springer.
Winer, B. J. (1971). *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill.

# Applied Research

Applied research is inquiry using the application of scientific methodology with the purpose of generating empirical observations to solve critical problems in society. It is widely used in varying contexts, ranging from applied behavior analysis to city planning and public policy and to program evaluation. Applied research can be executed through a diverse range of research strategies that can be solely quantitative, solely qualitative, or a mixed method research design that combines quantitative and qualitative data slices in the same project. What all the multiple facets in applied research projects share is one basic commonality—the practice of conducting research in "nonpure" research conditions because data are needed to help solve a real-life problem.

The most common way applied research is understood is by comparing it to *basic research*. Basic research—"pure" science—is grounded in the scientific method and focuses on the production of new knowledge and is not expected to have an immediate practical application. Although the distinctions between the two contexts are arguably somewhat artificial, researchers commonly identify four differences between applied research and basic research. Applied research differs from basic research in terms of purpose, context, validity, and methods (design).

## Research Purpose

The purpose of applied research is to increase what is known about a problem with the goal of creating a better solution. This is in contrast to basic research, in which the primary purpose is to expand on what is known—knowledge—with little significant connections to contemporary problems. A simple contrast that shows how research purpose differentiates these two lines of investigation can be seen in applied behavior analysis and psychological research. Applied behavior is a branch of psychology that generates empirical observations that focus at the level of the individual with the goal of developing effective interventions to solve specific problems. Psychology, on the other hand, conducts research to test theories or explain changing trends in certain populations.

The irrelevance of basic research to immediate problems may at times be overstated. In one form or another, observations generated in basic research eventually influence what we know about contemporary problems. Going back to the previous comparison, applied behavior investigators commonly integrate findings generated by cognitive psychologists—how people organize and analyze information—in explaining specific types of behaviors and identifying relevant courses of interventions to modify them. The question is, how much time needs to pass (5 months, 5 years, 50 years) in the practical application of research results in order for the research to be deemed basic research? In general, applied research observations are intended to be implemented in the first few years whereas basic researchers make no attempt to identify when their observations will be realized in everyday life.

## Research Context

The point of origin at which a research project begins is commonly seen as the most significant difference between applied research and basic research. In applied research, the context of pressing issues marks the beginning in a line of investigation. Applied research usually begins when a client has a need for research to help solve a problem. The context the client operates in provides the direction the applied investigator takes in terms of developing the research questions. The client usually takes a commanding role in framing applied research questions. Applied research questions tend to be open ended because the client sees the investigation as being part of a larger context made up of multiple stakeholders who understand the problem from various perspectives.

Basic research begins with a research question that is grounded in theory or previous empirical investigations. The context driving basic research takes one of two paths: testing the accuracy of hypothesized relationships among identified variables or confirming existing knowledge from earlier studies. In both scenarios, the basic research investigator usually initiates the research project based on his or her ability to isolate observable variables and to control and monitor the environment in which they operate. Basic research questions are narrowly defined and are investigated with only one level of analysis: prove or disprove theory or confirm or not confirm earlier research conclusions.

The contrast in the different contexts between applied research and basic research is simply put by Jon S. Bailey and Mary R. Burch in their explanation of applied behavior research in relation to psychology. The contrast can be pictured like this: In applied behavior research, subjects walk in the door with unique family histories that are embedded in distinct communities. In basic research, subjects "come in packing crates from a breeding farm, the measurement equipment is readily available, the experimental protocols are already established, and the research questions are derivative" (p. 3).

## Emphasis on Validity

The value of all research—applied and basic—is determined by its ability to address questions of internal and external validity. Questions of *internal validity* ask whether the investigator makes the correct observation on the causal relationship among identified variables. Questions of *external validity* ask whether the investigators appropriately generalize observations from their research project to relevant situations. A recognized distinction is that applied research values external validity more than basic research projects do. Assuming an applied research project adequately addresses questions of internal validity, its research conclusions are more closely assessed in how well they apply directly to solving problems.

Questions of internal validity play a more significant role in basic research. Basic research focuses on capturing, recording, and measuring causal relationships among identified variables. The application of basic research conclusions focuses more on their relevance to theory and the advancement of knowledge than on their generalizability to similar situations.

The difference between transportation planning and transportation engineering is one example of the different validity emphasis in applied research and basic research. Transportation planning is an applied research approach that is concerned with the siting of streets, highways, sidewalks, and public transportation to facilitate the efficient movement of goods and people. Transportation planning research is valued for its ability to answer questions of external validity and address transportation needs and solve traffic problems, such as congestion at a specific intersection. Traffic engineering is the basic research approach to studying function, design, and operation of transportation facilities and looks at the interrelationship of variables that create conditions for the inefficient movement of goods and people. Traffic engineering is valued more for its ability to answer questions of internal validity in correctly identifying the relationship among variables that can cause traffic and makes little attempt to solve specific traffic problems.

## Research Design

Applied research projects are more likely follow a *triangulation* research design than are basic research investigations. Triangulation is the research strategy that uses a combination of multiple data sets, multiple investigators, multiple

theories, and multiple methodologies to answer research questions. This is largely because of the context that facilitates the need for applied research. Client-driven applied research projects tend to need research that analyzes a problem from multiple perspectives in order to address the many constituents that may be impacted by the study. In addition, if applied research takes place in a less than ideal research environment, multiple data sets may be necessary in order for the applied investigator to generate a critical mass of observations to be able to make defensible conclusions about the problem at hand.

Basic research commonly adheres to a single-method, single-data-research strategy. The narrow focus in basic research requires the investigator to eliminate possible research variability (bias) to better isolate and observe changes in the studied variables. Increasing the number of types of data sets accessed and methods used to obtain them increases the possible risk of contaminating the basic research laboratory of observations.

Research design in transportation planning is much more multifaceted than research design in traffic engineering. This can be seen in how each approach would go about researching transportation for older people. Transportation planners would design a research strategy that would look at the needs of a specific community and assess several different data sets (including talking to the community) obtained through several different research methods to identify the best combination of interventions to achieve a desired outcome. Traffic engineers will develop a singular research protocol that focuses on total population demand in comparison with supply to determine unmet transportation demand of older people.

*John Gaber*

***See also*** Planning Research; Scientific Method

**Further Readings**

Baily, S. J., & Burch, M. R. (2002). *Research methods in applied behavior analysis*. Thousand Oaks, CA: Sage.

Bickman, L., & Rog, D. J. (1998). *Handbook of applied social research methods*. Thousand Oaks, CA: Sage.

Kimmel, J. A. (1988). *Ethics and values in applied social research*. Newbury Park, CA: Sage.

## A Priori Monte Carlo Simulation

An a priori Monte Carlo simulation is a special case of a Monte Carlo simulation that is used in the design of a research study, generally when analytic methods do not exist for the goal of interest for the specified model or are not convenient. A Monte Carlo simulation is generally used to evaluate empirical properties of some quantitative method by generating random data from a population with known properties, fitting a particular model to the generated data, collecting relevant information of interest, and replicating the entire procedure a large number of times (e.g., 10,000). In an a priori Monte Carlo simulation study, interest is generally in the effect of design factors on the inferences that can be made rather than a general attempt at describing the empirical properties of some quantitative method. Three common categories of design factors used in a priori Monte Carlo simulations are sample size, model misspecification, and unsatisfactory data conditions. As with Monte Carlo methods in general, the computational tediousness of a priori Monte Carlo simulation methods essentially requires one or more computers because of the large number of replications and thus the heavy computational load. Computational loads can be very great when the a priori Monte Carlo simulation is implemented for methods that are themselves computationally tedious (e.g., bootstrap, multilevel models, and Markov chain Monte Carlo methods).

For an example of when an a priori Monte Carlo simulation study would be useful, Ken Kelley and Scott Maxwell have discussed sample size planning for multiple regression when interest is in sufficiently narrow confidence intervals for standardized regression coefficients (i.e., the accuracy-in-parameter-estimation approach to sample size planning). Confidence intervals based on noncentral $t$ distributions should be used for standardized regression coefficients. Currently, there is no analytic way to plan for the sample size so that the computed interval will be no larger than desired some specified percent of the time. However, Kelley and Maxwell suggested an a priori Monte Carlo simulation

procedure when random data from the situation of interest are generated and a systematic search (e.g., a sequence) of different sample sizes is used until the minimum sample size is found at which the specified goal is satisfied.

As another example of when an application of a Monte Carlo simulation study would be useful, Linda Muthén and Bengt Muthén have discussed a general approach to planning appropriate sample size in a confirmatory factor analysis and structural equation modeling context by using an a priori Monte Carlo simulation study. In addition to models in which all the assumptions are satisfied, Muthén and Muthén suggested sample size planning using a priori Monte Carlo simulation methods when data are missing and when data are not normal—two conditions most sample size planning methods do not address.

Even when analytic methods do exist for designing studies, sensitivity analyses can be implemented within an a priori Monte Carlo simulation framework. Sensitivity analyses in an a priori Monte Carlo simulation study allow the effect of misspecified parameters, misspecified models, and/or the validity of the assumptions on which the method is based to be evaluated. The generality of the a priori Monte Carlo simulation studies is its biggest advantage. As Maxwell, Kelley, and Joseph Rausch have stated, "Sample size can be planned for any research goal, on any statistical technique, in any situation with an a priori Monte Carlo simulation study" (2008, p. 553).

*Ken Kelley*

***See also*** Accuracy in Parameter Estimation; Monte Carlo Simulation; Power Analysis; Sample Size Planning

**Further Readings**

Kelley, K. (2007). Sample size planning for the coefficient of variation: Accuracy in parameter estimation via narrow confidence intervals. *Behavior Research Methods, 39*(4), 755–766.

Kelley, K., & Maxwell, S. E. (2008). Power and accuracy for omnibus and targeted effects: Issues of sample size planning with applications to multiple regression. In P. Alasuuta, L. Bickman, & J. Brannen (Eds.), *The SAGE handbook of social research methods* (pp. 166–192). Thousand Oaks, CA: Sage.

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology, 59,* 537–563.

Muthén, L., & Muthén, B. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling, 4,* 599–620.

# *Aptitudes and Instructional Methods*

Research on the interaction between student characteristics and instructional methods is important because it is commonly assumed that different students learn in different ways. That assumption is best studied by investigating the interaction between student characteristics and different instructional methods. The study of that interaction received its greatest impetus with the publication of Lee Cronbach and Richard Snow's *Aptitudes and Instructional Methods* in 1977, which summarized research on the interaction between aptitudes and instructional treatments, subsequently abbreviated as ATI research. Cronbach and Snow indicated that the term *aptitude,* rather than referring exclusively to cognitive constructs, as had previously been the case, was intended to refer to any student characteristic. Cronbach stimulated research in this area in earlier publications suggesting that ATI research was an ideal meeting point between the usually distinct research traditions of correlational and experimental psychology. Before the 1977 publication of *Aptitudes and Instructional Methods,* ATI research was spurred by Cronbach and Snow's technical report summarizing the results of such studies, which was expanded in 1977 with the publication of the volume.

## Background

When asked about the effectiveness of different treatments, educational researchers often respond that "it depends" on the type of student exposed to the treatment, implying that the treatment interacted with some student characteristic. Two types of interactions are important in ATI research: *ordinal* and *disordinal,* as shown in Figure 1. In ordinal interactions (top two lines in Figure 1), one treatment yields superior outcomes at all levels of the student characteristic, though the difference between the

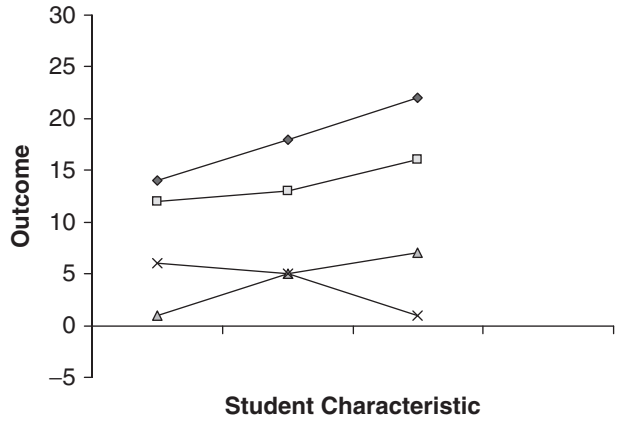



**Figure 1** Ordinal and Disordinal Interactions

outcomes is greater at one part of the distribution than elsewhere. In disordinal interactions (the bottom two lines in Figure 1), one treatment is superior at one point of the student distribution while the other treatment is superior for students falling at another point. The slope difference in ordinal interactions indicates that ultimately they are also likely to be disordinal, that is, the lines will cross at a further point of the student characteristic distribution than observed in the present sample.

## Research Design

ATI studies typically provide a segment of instruction by two or more instructional methods that are expected to be optimal for students with different characteristics. Ideally, research findings or some strong theoretical basis should exist that leads to expectations of differential effectiveness of the instruction for students with different characteristics. Assignment to instructional method may be entirely random or random within categories of the student characteristic. For example, students may be randomly assigned to a set of instructional methods and their anxiety then determined by some measure or experimental procedure. Or, in quasi-experimental designs, high- and low-anxiety students may be determined first and then—within the high- and low-anxiety groups—assignment to instructional methods should be random.

ATI research was traditionally analyzed with analysis of variance (ANOVA). The simplest ATI design conforms to a 2 × 2 ANOVA, with two treatment groups and two groups (high and low) on the student characteristic. In such studies, main effects were not necessarily expected for either the treatment or the student characteristic, but the interaction between them is the result of greatest interest.

Cronbach and Snow pointed out that in ANOVA designs, the student characteristic examined was usually available as a continuous score that had at least ordinal characteristics, and the research groups were developed by splitting the student characteristic distribution at some point to create groups (high and low; high, medium, and low; etc.). Such division into groups ignored student differences within each group and reduced the available variance by an estimated 34%. Cronbach and Snow recommended that research employ multiple linear regression analysis in which the treatments would be represented by so-called dummy variables and the student characteristic could be analyzed as a continuous score. It should also be noted, however, that when the research sample is at extreme ends of the distribution (e.g., one standard deviation above or below the mean), the use of ANOVA maximizes the possibility of finding differences between the groups.

## ATI Research Review

Reviews of ATI research reported few replicated interactions. Among the many reasons for these inconsistent findings were vague descriptions of the instructional treatments and sketchy relationships between the student characteristic and the instruction. Perhaps the most fundamental reason for the inconsistent findings was the inability to identify the cognitive processes required by the instructional treatments and engaged by the student characteristic. Slava Kalyuga, Paul Ayres, Paul Chandler, and John Sweller demonstrated that when the cognitive processes involved in instruction have been clarified, more consistent ATI findings have been reported and replicated.

Later reviews of ATI research, such as those by J. E. Gustaffson and J. O. Undheim or by Sigmund Tobias, reported consistent findings for Tobias's general hypothesis that students with limited knowledge of a domain needed instructional

support, that is, assistance to the learner, whereas more knowledgeable students could succeed without it. The greater consistency of interactions involving prior knowledge as the student characteristic may be attributable to some attributes of such knowledge. Unlike other student characteristics, prior domain knowledge contains the cognitive processes to be used in the learning of that material. In addition, the prior knowledge measure is likely to have been obtained in a situation fairly similar to the one present during instruction, thus also contributing any variance attributable to *situativity* to the results.

*Sigmund Tobias*

***See also*** Analysis of Covariance (ANCOVA); Interaction; Reactive Arrangements

**Further Readings**

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12,* 671–684.

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist, 30,* 116–127.

Cronbach, L. J., & Snow, R. E. (1969). *Individual differences and learning ability as a function of instructional variables.* (Contract No. OEC 4–6-061269–1217). Stanford, CA: Stanford University, School of Education.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods.* New York: Irvington Press.

Gustaffson, J., & Undheim, J. O. (1996). Individual differences in cognitive functions. In D. S. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 186–242). New York: Macmillan Reference.

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist, 38,* 23–31.

Tobias, S. (1976). Achievement treatment interactions. *Review of Educational Research, 46,* 61–74.

Tobias, S. (1982). When do instructional methods make a difference? *Educational Researcher, 11*(4), 4–9.

Tobias, S. (1989). Another look at research on the adaptation of instruction to student characteristics. *Educational Psychologist, 24,* 213–227.

Tobias, S. (2009). An eclectic appraisal of the success or failure of constructivist instruction: In S. Tobias & T. D. Duffy (Eds.), *Constructivist theory applied to education: Success or failure?* (pp. 335–350). New York: Routledge.

## Aptitude-Treatment Interaction

There are countless illustrations in the social sciences of a description of a phenomenon existing for many years before it is labeled and systematized as a scientific concept. One such example is in Book II of Homer's *Iliad,* which presents an interesting account of the influence exerted by Agamemnon, king of Argos and commander of the Greeks in the Trojan War, on his army. In particular, Homer describes the behavior of Odysseus, a legendary king of Ithaca, and the behavior of Thersites, a commoner and rank-and-file soldier, as contrasting responses to Agamemnon's leadership and role as "the shepherd of the people." Odysseus, Homer says, is "brilliant," having "done excellent things by thousands," while he describes Thersites as that "who knew within his head many words, but disorderly," and "this thrower of words, this braggart." Where the former admires the leadership of Agamemnon, accepts his code of honor, and responds to his request to keep the sage of Troy, the latter accuses Agamemnon of greed and promiscuity and demands a return to Sparta.

The observation that an intervention—educational, training, therapeutic, or organizational—when delivered the same way to different people might result in differentiated outcomes, was made a long time ago, as long as the eighth century BCE, as exemplified by Homer. In attempts to comprehend and explain this observation, researchers and practitioners have focused primarily on the concept of individual differences, looking for main effects that are attributable to concepts such as ability, personality, motivation, or attitude. When these inquiries started, early in the 20th century, not many parallel interventions were available. In short, the assumption at the time was that a student (a trainee in a workplace, a client in a clinical setting, or a soldier on a battlefield) possessed specific characteristics, such as Charles Spearman's *g* factor of intelligence, that could predict his or her success or failure in a training situation. However, this attempt to explain the success of an intervention by the characteristics of the intervenee was challenged by the appearance of multiple parallel interventions aimed at arriving at the same desired goal by

employing various strategies and tactics. It turned out that there were no ubiquitous collections of individual characteristics that would always result in success in a situation. Moreover, as systems of intervention in education, work training in industry, and clinical fields developed, it became apparent that different interventions, although they might be focused on the same target (e.g., teaching children to read, training bank tellers to operate their stations, helping a client overcome depression, or preparing soldiers for combat), clearly worked differently for different people. It was then suggested that the presence of differential outcomes of the same intervention could be explained by aptitude-treatment interaction (ATI, sometimes also abbreviated as AxT), a concept that was introduced by Lee Cronbach in the second part of the 20th century.

ATI methodology was developed to coaccount both for the individual characteristics of the intervenee and the variations in the interventions while assessing the extent to which alternative forms of interventions might have differential outcomes as a function of the individual characteristics of the person to whom the intervention is being delivered. In other words, investigations of ATI have been designed to determine whether particular treatments can be selected or modified to optimally serve individuals possessing particular characteristics (i.e., ability, personality, motivation). Today, ATI is discussed in three different ways: as a concept, as a method for assessing interactions among person and situation variables, and as a framework for theories of aptitude and treatment.

## ATI as a Concept

ATI as a concept refers to both an outcome and a predictor of that outcome. Understanding these facets of ATI requires decomposing the holistic concept into its three components—treatment, aptitude, and the interaction between them. The term *treatment* is used to capture any type of manipulation aimed at changing something. Thus, with regard to ATI, treatment can refer to a specific educational intervention (e.g., the teaching of equivalent fractions) or conceptual pedagogical framework (e.g., Waldorf pedagogy), a particular training (e.g., job-related activity, such as mastering a new piece of equipment at a work place) or self-teaching (e.g., mastering a new skill such as typing), a clinical manipulation (e.g., a session of massage) or long-term therapy (e.g., psychoanalysis), or inspiring a soldier to fight a particular battle (e.g., issuing an order) or preparing troops to use new strategies of war (e.g., fighting insurgency). *Aptitude* is used to signify any systematic measurable dimension of individual differences (or a combination of such) that is related to a particular treatment outcome. In other words, aptitude does not necessarily mean a level of general cognitive ability or intelligence; it can capture specific personality traits or transient psychological states. The most frequently studied aptitudes of ATI are in the categories of cognition, conation, and affection, but aptitudes are not limited to these three categories. Finally, *interaction* demarcates the degree to which the results of two or more interventions will differ for people who differ in one or more aptitudes. Of note is that interaction here is defined statistically and that both intervention and aptitude can be captured by qualitative or quantitative variables (observed, measured, self-reported, or derived). Also of note is that, being a statistical concept, ATI behaves just as any statistical interaction does. Most important, it can be detected only when studies are adequately powered. Moreover, it acknowledges and requires the presence of main effects of the aptitude (it has to be a characteristic that matters for a particular outcome, e.g., general cognitive ability rather than shoe size for predicting a response to educational intervention) and the intervention (it has to be an effective treatment that is directly related to an outcome, e.g., teaching a concept rather than just giving students candy). This statistical aspect of ATI is important for differentiating it from what is referred to by the ATI developers and proponents as *transaction*. Transaction signifies the way in which ATI is constructed, the environment and the process in which ATI emerges; in other words, ATI is always a statistical result of a transaction through which a person possessing certain aptitudes experiences a certain treatment. ATI as an outcome identifies combinations of treatments and aptitudes that generate a significant change or a larger change compared with other combinations. ATI as a predictor points to which treatment or treatments are more likely to generate

significant or larger change for a particular individual or individuals.

## ATI as a Method

ATI as a method permits the use of multiple experimental designs. The very premise of ATI is its capacity to combine correlational approaches (i.e., studies of individual differences) and experimental approaches (i.e., studies of interventional manipulations). Multiple paradigms have been developed to study ATI; many of them have been and continue to be applied in other, non-ATI, areas of interventional research. In classical accounts of ATI, the following designs are typically mentioned. In a *simple standard randomized between-persons design,* the outcome is investigated for persons who score at different levels of a particular aptitude when multiple, distinct interventions are compared. Having registered these differential outcomes, intervention selection is then carried out based on a particular level of aptitude to optimize the outcome. Within this design, often, when ATI is registered, it is helpful to carry out additional studies (e.g., case studies) to investigate the reason for the manifestation of ATI. The *treatment revision design* assumes the continuous adjustment of an intervention (or the creation of multiple parallel versions of it) in response to how persons with different levels of aptitude react to each improvement in the intervention (or alternative versions of the intervention). The point here is to optimize the intervention by creating its multiple versions or its multiple stages so that the outcome is optimized at all levels of aptitude. This design has between- and within-person versions, depending on the purposes of the intervention that is being revised (e.g., ensuring that all children can learn equivalent fractions regardless of their level of aptitude or ensuring the success of the therapy regardless of the variability in depressive states of a client across multiple therapy sessions). In the *aptitude growth design,* the target of intervention is the level of aptitude. The idea here is that as the level of aptitude changes, different types of interventions might be used to optimize the outcome. This type of design is often used in combination with growth-curve analyses. It can be applied as either between-persons or within-person designs. Finally, a type of design that has been gaining much popularity lately is the *regression discontinuity design.* In this design, the presence of ATI is registered when the same intervention is administered before and after a particular event (e.g., a change in aptitude in response to linguistic immersion while living in a country while continuing to study the language of that country).

## ATI as a Theoretical Framework

ATI as a theoretical framework underscores the flexible and dynamic, rather than fixed and deterministic, nature of the coexistence (or coaction) of individual characteristics (i.e., aptitudes) and situations (i.e., interventions). As a theory, ATI captures the very nature of variation in learning—not everyone learns equally well from the same method of instruction, and not every method of teaching works for everyone; in training—people acquire skills in a variety of ways; in therapy—not everyone responds well to a particular therapeutic approach; and in organizational activities—not everyone prefers the same style of leadership. In this sense, as a theoretical framework, ATI appeals to professionals in multiple domains as it justifies the presence of variation in outcomes in classrooms, work environments, therapeutic settings, and battlefields. While applicable to all types and levels of aptitudes and all kinds of interventions, ATI is particularly aligned with more extreme levels of aptitudes, both low and high, and more specialized interventions. The theory of ATI acknowledges the presence of heterogeneity in both aptitudes and interventions, and its premise is to find the best possible combinations of the two to maximize the homogeneity of the outcome. A particular appeal of the theory is its transactional nature and its potential to explain and justify both success and failure in obtaining the desired outcome. As a theoretical framework, ATI does not require the interaction to either be registered empirically or be statistically significant. It calls for a theoretical examination of the aptitude and interventional parameters whose interaction would best explain the dynamics of learning, skill acquisition and demonstration, therapy, and leadership. The beneficiaries of this kind of examination are of two kinds. First, it is the researchers themselves. Initially thinking through experiments and field

studies before trying to confirm the existence of ATI empirically was, apparently, not a common feature of ATI studies during the height of their popularity. Perhaps a more careful consideration of the "what, how, and why" of measurement in ATI research would have prevented the observation that many ATI findings resulted from somewhat haphazard fishing expeditions, and the resulting views on ATI research would have been different. A second group of beneficiaries of ATI studies are practitioners and policy makers. That there is no intervention that works for all, and that one has to anticipate both successes and failures and consider who will and who will not benefit from a particular intervention, are important realizations to make while adopting a particular educational program, training package, therapeutic approach, or organizational strategy, rather than in the aftermath. However, the warning against embracing panaceas, made by Richard Snow, in interventional research and practice is still just a warning, not a common presupposition.

## Criticism

Having emerged in the 1950s, interest in ATI peaked in the 1970s and 1980s, but then dissipated. This expansion and contraction were driven by an initial surge in enthusiasm, followed by a wave of skepticism about the validity of ATI. Specifically, a large-scope search for ATI, whose presence was interpreted as being marked by differentiated regression slopes predicting outcomes from aptitudes for different interventions, or by the significance of the interaction terms in analysis of variance models, was enthusiastically carried out by a number of researchers. The accumulated data, however, were mixed and often contradictory—there were traces of ATI, but its presence and magnitude were not consistently identifiable or replicable. Many reasons have been mentioned in discussions of why ATI is so elusive: underpowered studies, weak theoretical conceptualizations of ATI, simplistic research designs, imperfections in statistical analyses, and the magnitude and even the nonexistence of ATI, among others. As a result of this discussion, the initial prediction of the originator of ATI's concept, Lee Cronbach, that interventions designed for the average individual would be ultimately replaced by multiple parallel interventions to fit groups of individuals, was revised. The "new view" of ATI, put forward by Cronbach in 1975, acknowledged that, although in existence, ATI is much more complex and fluid than initially predicted and ATI's dynamism and fluidity prevent professionals from cataloging specific types of ATI and generalizing guidelines for prescribing different interventions to people, given their aptitudes. Although the usefulness of ATI as a theory has been recognized, its features as a concept and as a method have been criticized along the lines of (a) our necessarily incomplete knowledge of all possible aptitudes and their levels, (b) the shortage of good psychometric instruments that can validly and reliably quantify aptitudes, (c) the biases inherent in many procedures related to aptitude assessment and intervention delivery, and (d) the lack of understanding and possible registering of important "other" nonstatistical interactions (e.g., between student and teacher, client and therapist, environment and intervention). And yet ATI has never been completely driven from the field, and there have been steady references to the importance of ATI's framework and the need for better-designed empirical studies of ATI.

## Gene × Environment Interaction

ATI has a number of neighboring concepts that also work within the general realm of qualifying and quantifying individual differences in situations of acquiring new knowledge or new skills. Among these concepts are learning styles, learning strategies, learning attitudes, and many interactive effects (e.g., aptitude-outcome interaction). Quite often, the concept of ATI is discussed side by side with these neighboring concepts. Of particular interest is the link between the concept of ATI and the concept of Gene × Environment interaction (G × E). The concept of G × E first appeared in nonhuman research but gained tremendous popularity in the psychological literature within the same decade. Of note is that the tradition of its use in this literature is very similar to that of the usage of ATI; specifically, G × E also can be viewed as a concept, a method, and a theoretical framework. But the congruence between the two concepts is incomplete, of course; the concept of G × E adopts a very narrow definition of aptitude, in which individual differences are reduced to

genetic variation, and a very broad definition of treatment, in which interventions can be equated with live events. Yet an appraisal of the parallels between the concepts of ATI and G × E is useful because it captures the field's desire to engage interaction effects for explanatory purposes whenever the explicatory power of main effects is disappointing. And it is interesting that the accumulation of the literature on G × E results in a set of concerns similar to those that interrupted the golden rush of ATI studies in the 1970s.

Yet methodological concerns aside, the concept of ATI rings a bell for all of us who have ever tried to learn anything in a group of people: what works for some of us will not work for the others as long as we differ on even one characteristic that is relevant to the outcome of interest. Whether it was wit or something else by which Homer attempted to differentiate Odysseus and Thersites, the poet did at least successfully make an observation that has been central to many fields of social studies and that has inspired the appearance of the concept, methodology, and theoretical framework of ATI, as well as the many other concepts that capture the essence of what it means to be an individual in any given situation: that individual differences in response to a common intervention exist. Millennia later, it is an observation that still claims our attention.

*Elena L. Grigorenko*

***See also*** Effect Size, Measures of; Field Study; Growth Curve; Interaction; Intervention; Power; Within-Subjects Design

**Further Readings**

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12,* 671–684.

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist, 30,* 116–127.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions*. Irvington, UK: Oxford University Press.

Dance, K. A., & Neufeld, R. W. J. (1988). Aptitude-treatment interaction research in the clinical setting: A review of attempts to dispel the "patient uniformity" myth. *Psychological Bulletin, 194,* 192–213.

Fuchs, L. S., & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children, 53,* 199–208.

Grigorenko, E. L. (2005). The inherent complexities of gene-environment interactions. *Journal of Gerontology, 60B,* 53–64.

Snow, R. E. (1984). Placing children in special education: Some comments. *Educational Researcher, 13,* 12–14.

Snow, R. E. (1991). Aptitude-treatment interaction as a framework for research on individual differences in psychotherapy. *Journal of Consulting & Clinical Psychology, 59,* 205–216.

Spearman, C. (1904). "General intelligence" objectively determined and measured. *American Journal of Psychology, 15,* 201–293.

Violato, C. (1988). Interactionism in psychology and education: A new paradigm or source of confusion? *Journal of Education Thought, 22,* 4–20.

# Assent

The term *assent* refers to the verbal or written agreement to engage in a research study. Assent is generally applicable to children between the ages of 8 and 18 years, although assent may apply to other vulnerable populations also.

Vulnerable populations are those composed of individuals who are unable to give consent due to diminished autonomy. Diminished autonomy occurs when an individual is incapacitated, has restricted freedom, or is a minor. Understanding the relevance of assent is important because without obtaining the assent of a participant, the researcher has restricted the freedom and autonomy of the participant and in turn has violated the basic ethical principle of respect for persons. Assent with regard to vulnerable populations is discussed here, along with the process of obtaining assent and the role of institutional review boards in the assent process.

## Vulnerable Populations

Respect for persons requires that participants agree to engage in research voluntarily and have adequate information to make an informed decision. Most laws recognize that a person 18 years of age or older is able to give his or her informed *consent* to participate in the research study. However, in

some cases individuals lack the capacity to provide informed consent. An individual may lack the capacity to give his or her consent for a variety of reasons; examples include a prisoner who is ordered to undergo an experimental treatment designed to decrease recidivism, a participant with mental retardation, or an older adult with dementia whose caretakers believe an experimental psychotherapy group may decrease his or her symptoms. Each of the participants in these examples is not capable of giving permission to participate in the research because he or she either is coerced into engaging in the research or lacks the ability to understand the basic information necessary to fully consent to the study.

State laws prohibit minors and incapacitated individuals from giving consent. In these cases, permission must be obtained from parents and court-appointed guardians, respectively. However, beyond consent, many ethicists, professional organizations, and ethical codes require that *assent* be obtained. With children, state laws define when a young person is legally competent to make informed decisions. Some argue that the ability to give assent is from 8 to 14 years of age because the person is able to comprehend the requirements of the research. In general, however, it is thought that by the age of 10, children should be able to provide assent to participate. It is argued that obtaining assent increases the autonomy of the individual. By obtaining assent, individuals are afforded as much control as possible over their decision to engage in the research given the circumstances, regardless of their mental capacity.

## Obtaining Assent

Assent is not a singular event. It is thought that assent is a continual process. Thus, researchers are encouraged to obtain permission to continue with the research during each new phase of research (e.g., moving from one type of task to the next). If an individual assents to participate in the study but during the study requests to discontinue, it is recommended that the research be discontinued.

Although obtaining assent is strongly recommended, failure to obtain assent does not necessarily preclude the participant from engaging in the research. For example, if the parent of a 4-year-old child gives permission for the child to attend a social skills group for socially anxious children, but the child does not assent to treatment, the child may be enrolled in the group without his or her assent. However, it is recommended that assent be obtained whenever possible. Further, if a child does not give assent initially, attempts to obtain assent should continue throughout the research. Guidelines also suggest that assent may be overlooked in cases in which the possible benefits of the research outweigh the costs. For example, if one wanted to study the effects of a life-saving drug for children and the child refused the medication, the benefit of saving the child's life outweighs the cost of not obtaining assent. Assent may be overlooked in cases in which assent of the participants is not feasible, as would be the case of a researcher interested in studying children who died as a result of not wearing a seatbelt.

Obtaining assent is an active process whereby the participant and the researcher discuss the requirements of the research. In this case, the participant is active in the decision making. *Passive consent,* a concept closely associated with assent and consent, is the lack of protest, objection, or opting out of the research study and is considered permission to continue with the research.

## Institutional Review Boards

Institutional review boards frequently make requirements as to the way assent is to be obtained and documented. Assent may be obtained either orally or in writing and should always be documented. In obtaining assent, the researcher provides the same information as is provided to an individual from whom consent is requested. The language level and details may be altered in order to meet the understanding of the assenting participant. Specifically, the participant should be informed of the purpose of the study; the time necessary to complete the study; as well as the risks, benefits, and alternatives to the study or treatment. Participants should also have access to the researcher's contact information. Finally, limits of confidentiality should be addressed. This is particularly important for individuals in the prison system and for children.

*Tracy J. Cohn*

*See also* Debriefing; Ethics in the Research Process; Informed Consent; Interviewing

### Further Readings

*Belmont report: Ethical principles and guidelines for the protection of human subjects of research*. (1979). Washington, DC: U.S. Government Printing Office.

Grisso, T. (1992). Minors' assent to behavioral research without parental consent. In B. Stanley & J. E. Sieber (Eds.), *Social research on children and adolescents: Ethical issues* (pp. 109–127). Newbury Park, CA: Sage.

Miller, V. A., & Nelson, R. M. (2006). A developmental approach to child assent for nontherapeutic research. *Journal of Pediatrics, 150*(4), 25–30.

Ross, L. F. (2003). Do healthy children deserve greater protection in medical research? *Journal of Pediatrics, 142*(2), 108–112.

## Association, Measures of

Measuring association between variables is very relevant for investigating *causality,* which is, in turn, the sine qua non of scientific research. However, an association between two variables does not necessarily imply a causal relationship, and the research design of a study aimed at investigating an association needs to be carefully considered in order for the study to obtain valid information. Knowledge of measures of association and the related ideas of *correlation, regression,* and causality are cornerstone concepts in research design. This entry is directed at researchers disposed to approach these concepts in a conceptual way.

### Measuring Association

In scientific research, association is generally defined as the statistical dependence between two or more variables. Two variables are associated if some of the variability of one variable can be accounted for by the other, that is, if a change in the quantity of one variable conditions a change in the other variable.

Before investigating and measuring association, it is first appropriate to identify the types of variables that are being compared (e.g., nominal, ordinal, discrete, continuous). The type of variable will determine the appropriate statistical technique or test that is needed to establish the existence of an association. If the statistical test shows a conclusive association that is unlikely to occur by random chance, different types of regression models can be used to quantify how change in *exposure* to a variable relates to the change in the *outcome* variable of interest.

#### *Examining Association Between Continuous Variables With Correlation Analyses*

Correlation is a measure of association between two variables that expresses the degree to which the two variables are rectilinearly related. If the data do not follow a straight line (e.g., they follow a curve), common correlation analyses are not appropriate. In correlation, unlike regression analysis, there are no dependent and independent variables.

When both variables are measured as discrete or continuous variables, it is common for researchers to examine the data for a correlation between these variables by using the *Pearson product-moment correlation coefficient* ($r$). This coefficient has a value between $-1$ and $+1$ and indicates the strength of the association between the two variables. A *perfect* correlation of $\pm 1$ occurs only when all pairs of values (or *points*) fall exactly on a straight line.

A positive correlation indicates in a broad way that increasing values of one variable correspond to increasing values in the other variable. A negative correlation indicates that increasing values in one variable corresponds to decreasing values in the other variable. A correlation value close to 0 means no association between the variables. The $r$ provides information about the strength of the correlation (i.e., the nearness of the points to a straight line). Figure 1 gives some examples of correlations, correlation coefficients, and related regression lines.

A condition for estimating correlations is that both variables must be obtained by random sampling from the same population. For example, one can study the correlation between height and weight in a sample of children but not the correlation between height and three different types of diet that have been decided by the investigator. In

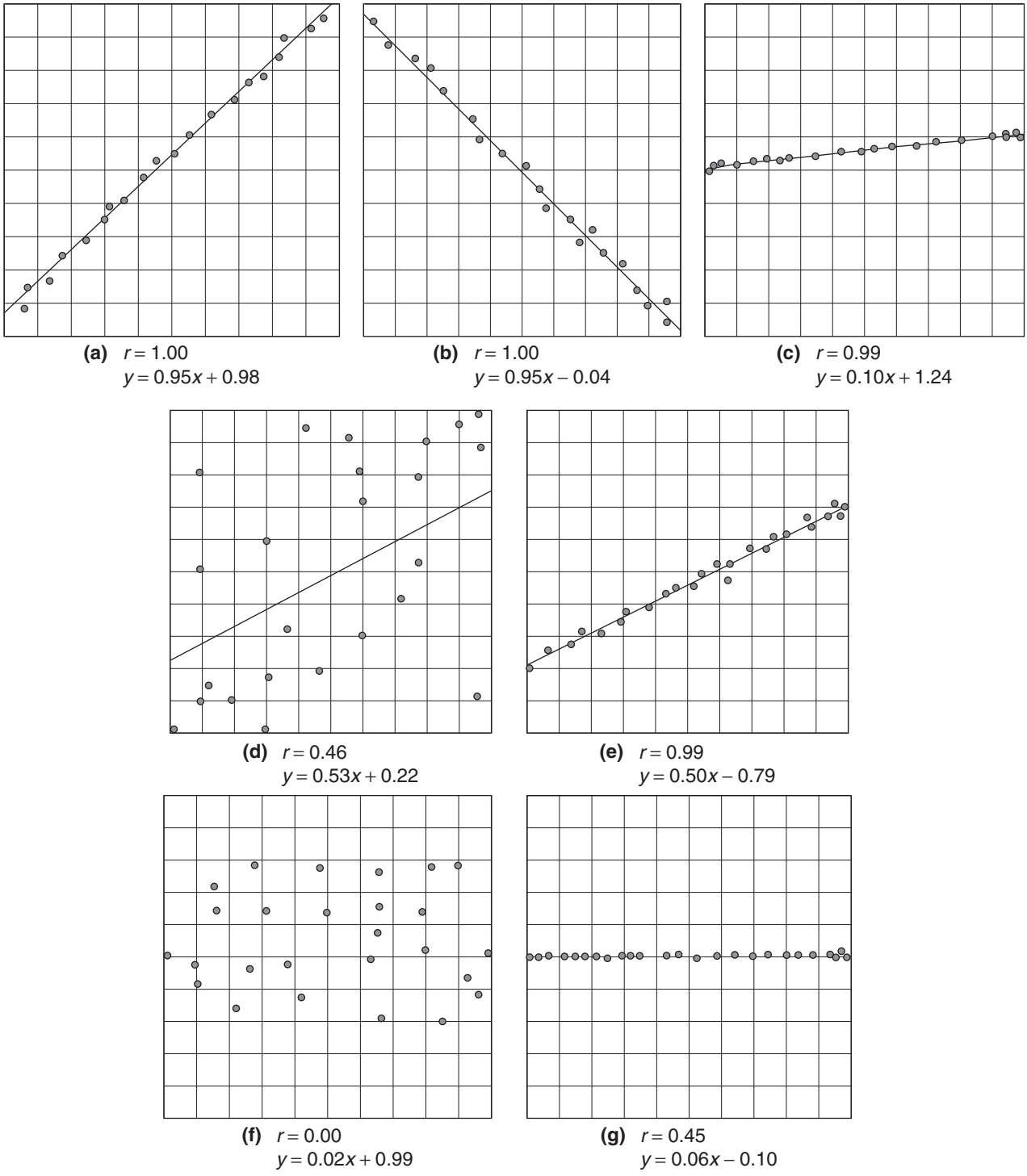



**Figure 1** Scatterplots for Correlations of Various Magnitudes

*Notes:* Simulation examples show perfect positive (a, c) and negative (b) correlations, as well as regression lines with similar correlation coefficients but different slopes (a, c). The figure also shows regression lines with similar slopes but different correlation coefficients (d, e and f, g).

the latter case, it would be more appropriate to apply a regression analysis.

The Pearson correlation coefficient may not be appropriate if there are outliers (i.e., extreme values). Therefore, the first step when one is studying correlations is to draw a scatterplot of the two variables to examine whether there are any outliers. These variables should be standardized to the same scale before they are plotted.

If outliers are present, *nonparametric types of correlation coefficients* can be calculated to examine the linear association. The *Spearman rank correlation coefficient,* for example, calculates correlation coefficients based on the ranks of both variables. *Kendall's coefficient of concordance* calculates the concordance and discordance of the observed (or ranked) exposure and outcome variables between pairs of individuals.

When the variables one is investigating are nominal or have few categories or when the scatterplot of the variables suggests an association that is not rectilinear but, for example, quadratic or cubic, then the correlation coefficients des cribed above are not suitable. In these cases other approaches are needed to investigate the association between variables.

### *Chi-Square Tests of Association for Categorical Variables*

A common method for investigating "general" association between two categorical variables is to perform a *chi-square test*. This method compares the observed number of individuals within cells of a cross-tabulation of the categorical variables with the number of individuals one would expect in the cells if there was no association and the individuals were randomly distributed. If the observed and expected frequencies differ statistically (beyond random chance according to the chi-square distribution), the variables are said to be associated.

A chi-square *test for trend* can also examine for linear association when the exposure category is ordinal. Other statistical tests of association include measurements of agreement in the association, such as the kappa statistic or McNemar's test, which are suitable when the study design is a matched case-control.

### *Quantifying Association in General and by Regression Analyses in Particular*

In *descriptive* research, the occurrence of an outcome variable is typically expressed by group measurements such as averages, proportions, incidence, or prevalence rates. In *analytical* research, an association can be quantified by comparing, for example, the absolute risk of the outcome in the exposed group and in the nonexposed group. Measurements of association can then be expressed either as *differences* (difference in risk) or as *ratios,* such as relative risks (a ratio of risks) or odds ratios (a ratio of odds), and so forth. A ratio with a numerical value greater than 1 (greater than 0 for differences) indicates a positive association between the exposure variable and the outcome variables, whereas a value less than 1 (less than 0 for differences) indicates a negative association. These measures of association can be calculated from cross-tabulation of the outcome variable and exposure categories, or they can be estimated in regression models.

General measures of association such as correlation coefficients and chi-square tests are rather unspecific and provide information only on the existence and strength of an association. *Regression analysis,* however, attempts to model the relationship between two variables by fitting a linear equation to observed data in order to quantify, and thereby predict, the change in the outcome of interest with a unit increase in the exposure variable. In regression analysis, one variable is considered to be an *explanatory variable* (the exposure), and the other is considered to be a *dependent variable* (the outcome).

The method of least squares is the method applied most frequently for fitting a regression line. This method calculates the best-fitting line for the observed data by minimizing the sum of the squares of the vertical deviations from each data point to the line. When a point is placed exactly on the fitted line, its vertical deviation is 0. Deviations are also known as *residuals* or *errors*. The better an explanatory variable predicts the outcome, the lower is the sum of the squared residuals (i.e., residual variance).

A simple *linear regression model,* for example, can examine the increase in blood pressure with a unit increase in age with the regression model

$$Y = a + bX + e,$$

where $X$ is the explanatory variable (i.e., age in years) and $Y$ is the dependent variable (i.e., blood pressure in mm Hg). The slope of the line is $b$ and represents the change in blood pressure for every year of age. Observe that $b$ does not provide information about the strength of the association but only on the average change in $Y$ when $X$ increases by one unit. The strength of the association is indicated by the correlation coefficient $r$, which informs on the closeness of the points to the regression line (see Figure 1). The parameter $a$ is the intercept (the value of $y$ when $x = 0$), which corresponds to the mean blood pressure in the sample. Finally, $e$ is the residual or error.

A useful measure is the square of the correlation coefficient, $r^2$, also called the *coefficient of determination,* and it indicates how much of the variance in the outcome (e.g., blood pressure) is explained by the exposure (e.g., age). As shown in Figure 1, different $b$s can be found with similar $r$s, and similar $b$s can be observed with different $r$s. For example, many biological variables have been proposed as risk factors for cardiovascular diseases because they showed a high $b$ value, but they have been rejected as common risk factors because their $r^2$ was very low.

Different types of regression techniques are suitable for different outcome variables. For example, a *logistic regression* is suitable when the outcome is binary (i.e., 0 or 1), and logistic regression can examine, for example, the increased probability (more properly, the increase in log odds) of myocardial infarction with unit increase in age. The *multinomial regression* can be used for analyzing outcome with several categories. A *Poisson regression* can examine how rate of disease changes with exposure, and a *Cox regression* is suitable for survival analysis.

## Association Versus Causality

Exposure variables that show a statistical relationship with an outcome variable are said to be *associated* with the outcome. It is only when there is strong evidence that this association is causal that the exposure variable is said to *determine* the outcome. In everyday scientific work, researchers apply a pragmatic rather than a philosophical framework to identify causality. For example, researchers want to discover modifiable causes of a disease.

Statistical associations say nothing by themselves on causality. Their causal value depends on the knowledge background of the investigator and the research design in which these statistical associations are observed. In fact, the only way to be completely sure that an association is causal would be to observe the very same individual living two parallel and exactly similar lives except that in one life, the individual was exposed to the variable of interest, and in the other life, the same individual was not exposed to the variable of interest (a situation called *counterfactual*). In this *ideal design,* the two (hypothetical) parallel lives of the individual are exchangeable in every way except for the exposure itself.

While the ideal research design is a just a chimera, there are alternative approaches that try to approximate the ideal design by comparing similar groups of people rather than the same individual. One can, for example, perform an experiment by taking random samples from the same population and randomly allocating the exposure of interest to the samples (i.e., *randomized trials*). In a randomized trial, an association between the average level of exposure and the outcome is possibly causal. In fact, random samples from the same population are—with some random uncertainty—identical concerning both *measured* and *unmeasured* variables, and the random allocation of the exposure creates a counterfactual situation very appropriate for investigating causal associations. The randomized trial design is theoretically closest to the ideal design, but sometimes it is unattainable or unethical to apply this design in the real world.

If conducting a randomized trial is not possible, one can use *observational* designs in order to simulate the ideal design, at least with regard to measured variables. Among observational approaches it is common to use *stratification, restriction,* and *multiple regression* techniques. One can also take into account the propensity for exposure when comparing individuals or groups (e.g., *propensity scores techniques*), or one may investigate the same individual at two different times (*case crossover design*). On some occasions one may have access to *natural experiments* or *instrumental variables.*

When planning a research design, it is always preferable to perform a *prospective* study because it identifies the exposure before any individual has developed the outcome. If one observes an association in a *cross-sectional* design, one can never be sure of the direction of the association. For example, low income is associated with impaired health in cross-sectional studies, but it is not known whether bad health leads to low income or the opposite. As noted by Austin Bradford Hill, the existence of a *temporal relationship* is the main criterion for distinguishing causality from association. Other relevant criteria pointed out by this author are *consistency, strength, specificity, dose-response relationship, biological plausibility,* and *coherence*.

## Bias and Random Error in Association Studies

When planning a study design for investigating causal associations, one needs to consider the possible existence of *random error, selection bias, information bias,* and *confounding,* as well as the presence of *interactions* or *effect modification* and of *mediator* variables.

Bias is often defined as the lack of *internal validity* of the association between exposure and outcome variable of interest. This is in contrast to *external validity,* which concerns generalizability of the association to other populations. Bias can also be defined as nonrandom or *systematic* difference between an estimate and the true value of the population.

### *Random Error*

When designing a study, one always needs to include a sufficient number of individuals in the analyses to achieve appropriate *statistical power* and ensure that *conclusive estimates* of association can be obtained. Suitable statistical power is especially relevant when it comes to establishing the *absence* of association between two variables. Moreover, when a study involves a large number of individuals, more information is available. More information lowers the random error, which in turn increases the *precision* of the estimates.

### *Selection Bias*

Selection bias can occur if the sample differs from the rest of the population and if the observed association is *modified* by a third variable. The study sample may be different from the rest of the population (e.g., only men or only healthy people), but this situation does not necessarily convey that the results obtained are biased and cannot be applied to the general population. Many randomized clinical trials are performed on a restricted sample of individuals, but the results are actually generalizable to the whole population. However, if there is an *interaction* between variables, the *effect modification* that this interaction produces must be considered. For example, the association between exposure to asbestos and lung cancer is much more intense among smokers than among nonsmokers. Therefore, a study on a population of nonsmokers would not be generalizable to the general population. Failure to consider interactions may even render associations spurious in a sample that includes the whole population. For example, a drug may increase the risk of death in a group of patients but decrease this risk in other different groups of patients. However, an overall measure would show no association since the antagonistic directions of the underlying associations compensate each other.

### *Information Bias*

Information bias simply arises because information collected on the variables is erroneous. All variables must be measured correctly; otherwise, one can arrive at imprecise or even spurious associations.

### *Confounding*

An association between two variables can be confounded by a third variable. Imagine, for example, that one observes an association between the existence of yellow nails and mortality. The causality of this association could be plausible. Since nail tissue stores body substances, the yellow coloration might indicate poisoning or metabolic disease that causes an increased mortality. However, further investigation would indicate that individuals with yellow nails were actually heavy smokers. The habit of holding the cigarette between the fingers discolored their nails, but the cause of death was smoking. That is, smoking was associated with both yellow nails and mortality and originated a confounded association (Figure 2).

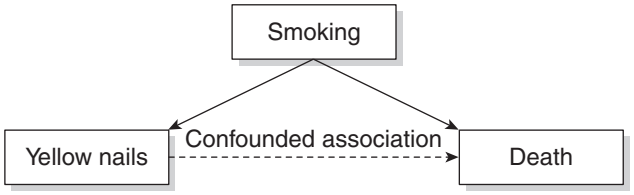



**Figure 2** Deceptive Correlation Between Yellow Nails and Mortality

*Note:* Because smoking is associated with both yellow nails and mortality, it originated a confounded association between yellow nails and mortality

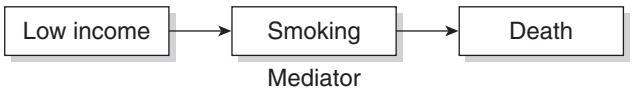



**Figure 3** Smoking Acts as a Mediator Between Income and Early Death

*Note:* Heavy smoking mediated the effect of low income on mortality.

### *Mediation*

In some cases an observed association is mediated by an *intermediate* variable. For example, individuals with low income present a higher risk of early death than do individuals with high income. Simultaneously, there are many more heavy smokers among people with low income. In this case, heavy smoking mediates the effect of low income on mortality.

Distinguishing which variables are confounders and which are mediators cannot be done by statistical techniques only. It requires previous knowledge, and in some cases variables can be both confounders and mediators.

### Directed Acyclic Graphs

Determining which variables are confounders, intermediates, or independently associated variables can be difficult when many variables are involved. Directed acyclic graphs use a set of simple rules to create a visual representation of direct and indirect associations of covariates and exposure variables with the outcome. These graphs can help researchers understand possible causal relationships.

*Juan Merlo and Kristian Lynch*

***See also*** Bias; Cause and Effect; Chi-Square Test; Confounding; Correlation; Interaction; Multiple Regression; Power Analysis

### Further Readings

Altman, D. G. (1991). *Practical statistics for medical research*. New York: Chapman & Hall/CRC.

Hernan, M. A., Hernandez-Diaz, S., Werler, M. M., & Mitchell, A. A. (2002). Causal knowledge as a prerequisite for confounding evaluation: An application to birth defects epidemiology. *American Journal of Epidemiology, 155*(2), 176–184.

Hernan, M. A., & Robins, J. M. (2006). Instruments for causal inference: An epidemiologist's dream? *Epidemiology, 17*(4), 360–372.

Hill, A. B. (1965). Environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine, 58,* 295–300.

Jaakkola, J. J. (2003). Case-crossover design in air pollution epidemiology. *European Respiratory Journal, 40,* 81s-85s.

Last, J. M. (Ed.). (2000). *A dictionary of epidemiology* (4th ed.). New York: Oxford University Press.

Liebetrau, A. M. (1983). *Measures of association.* Newbury Park, CA: Sage.

Lloyd, F. D., & Van Belle, G. (1993). *Biostatistics: A methodology for the health sciences*. New York: Wiley.

Oakes, J. M., & Kaufman J. S. (Eds.). (2006). *Methods in social epidemiology.* New York: Wiley.

Rothman, K. J. (Ed.). (1988). *Causal inference.* Chestnut Hill, MA: Epidemiology Resources.

Susser, M. (1991). What is a cause and how do we know one? A grammar for pragmatic epidemiology. *American Journal of Epidemiology, 33,* 635–648.

## AUTOCORRELATION

Autocorrelation describes sample *or* population observations or elements that are related to each other across time, space, or other dimensions. Correlated observations are common but problematic, largely because they violate a basic statistical assumption about many samples: independence across elements. Conventional tests of statistical significance assume simple random sampling, in which not only each element has an equal chance of selection but also each combination of elements has an equal chance of selection; autocorrelation

violates this assumption. This entry describes common sources of autocorrelation, the problems it can cause, and selected diagnostics and solutions.

## Sources

What is the best predictor of a student's 11th-grade academic performance? His or her 10th-grade grade point average. What is the best predictor of this year's crude divorce rate? Usually last year's divorce rate. The old slogan "Birds of a feather flock together" describes a college classroom in which students are about the same age, at the same academic stage, and often in the same disciplinary major. That slogan also describes many residential city blocks, where adult inhabitants have comparable incomes and perhaps even similar marital and parental status. When examining the spread of a disease, such as the H1N1 influenza, researchers often use epidemiological maps showing concentric circles around the initial outbreak locations.

All these are examples of correlated observations, that is, autocorrelation, in which two individuals from a classroom or neighborhood cluster, cases from a time series of measures, or proximity to a contagious event resemble each other more than two cases drawn from the total population of elements by means of a simple random sample. Correlated observations occur for several reasons:

- Repeated, comparable measures are taken on the same individuals over time, such as many pretest and posttest experimental measures or panel surveys, which reinterview the same individual. Because people remember their prior responses or behaviors, because many behaviors are habitual, and because many traits or talents stay relatively constant over time, these repeated measures become correlated for the same person.

- Time-series measures also apply to larger units, such as birth, divorce, or labor force participation rates in countries or achievement grades in a county school system. Observations on the same variable are repeated on the same unit at some periodic interval (e.g., annual rate of felony crimes). The units transcend the individual, and the periodicity of measurement is usually regular. A lag describes a measure of the same variable on the same unit at an earlier time, frequently one period removed (often called $t - 1$).

- Spatial correlation occurs in cluster samples (e.g., classrooms or neighborhoods): Physically adjacent elements have a higher chance of entering the sample than do other elements. These adjacent elements are typically more similar to already sampled cases than are elements from a simple random sample of the same size.

- A variation of spatial correlation occurs with contagion effects, such as crime incidence (burglars ignore city limits in plundering wealthy neighborhoods) or an outbreak of disease.

- Multiple (repeated) measures administered to the same individual at approximately the same time (e.g., a lengthy survey questionnaire with many Likert-type items in *agree-disagree* format).

## Autocorrelation Terms

The terms *positive* or *negative autocorrelation* often apply to time-series data. Societal inertia can inflate the correlation of observed measures across time. The social forces creating trends such as falling marriage rates or rising gross domestic product often carry over from one period into the next. When trends continue over time (e.g., a student's grades), positive predictions can be made from one period to the next, hence the term *positive autocorrelation.*

However, forces at one time can also create compensatory or corrective mechanisms at the next, such as consumers' alternating patterns of "save, then spend" or regulation of production based on estimates of prior inventory. The data points seem to ricochet from one time to the next, so adjacent observations are said to be negatively correlated, creating a cobweb-pattern effect.

The order of the autocorrelation process references the degree of periodicity in correlated observations. When adjacent observations are correlated, the process is *first-order autoregression,* or AR (1). If every other observation, or alternate observations, is correlated, this is an AR (2) process. If every third observation is correlated, this is an AR (3) process, and so on. The order of the process is important, first because the most

available diagnostic tests and corrections are for the simplest situation, an AR (1); higher order processes require more complex corrections. Second, the closer two observations are in time or space, the larger the correlation between them, creating more problems for the data analyst. An AR (1) process describes many types of autocorrelation, such as trend data or contagion effects.

## Problems

Because the similarity among study elements is more pronounced than that produced by other probability samples, each autocorrelated case "counts less" than a case drawn using simple random sampling. Thus, the "real" or corrected sample size when autocorrelation is present is smaller than a simple random sample containing the same number of elements. This statistical attenuation of the casebase is sometimes called the *design effect,* and it is well known to survey statisticians who design cluster samples.

The sample size is critical in inferential statistics. The $N$ comprises part of the formula for estimates of sample variances and the standard error. The standard error forms the denominator for statistics such as $t$ tests. The $N$ is also used to calculate degrees of freedom for many statistics, such as $F$ tests in analysis of variance or multiple regression, and it influences the size of chi-square.

When autocorrelation is present, use of the observed $N$ means overestimating the effective $n$. The calculated variances and standard errors that use simple random sampling formulae (as most statistics computer programs do) are, in fact, too low. In turn, this means $t$ tests and other inferential statistics are too large, leading the analyst to reject the null hypothesis inappropriately. In short, autocorrelation often leads researchers to think that many study results are statistically significant when they are not.

For example, in ordinary least squares (OLS) regression or analysis of variance, autocorrelation renders the simple random sampling formulae invalid for the error terms and measures derived from them. The true sum of squared errors ($\sigma$) is now inflated (often considerably) because it is divided by a fraction:

$$\sigma = \Sigma v^2/(1 - \rho^2),$$

where $v$ is the random component in a residual or error term.

### *Rho*

When elements are correlated, a systematic bias thus enters into estimates of the residuals or error terms. This bias is usually estimated numerically by rho ($\rho$), the intraclass correlation coefficient, or the correlation of autocorrelation. Rho estimates the average correlation among (usually) adjacent pairs of elements. Rho is found, sometimes unobtrusively, in many statistics that attempt to correct and compensate for autocorrelation, such as hierarchical linear models.

### *An Example: Autocorrelation Effects on Basic Regression Models*

With more complex statistical techniques, such as regression, the effects of $\rho$ multiply beyond providing a less stable estimate of the population mean. If autocorrelation occurs for scores on the dependent variable in OLS regression, then the regression residuals will also be autocorrelated, creating a systematic bias in estimates of the residuals and statistics derived from them. For example, standard computer OLS regression output will be invalid for the following: the residual sum of squares, the standard error of the (regression) estimate, the $F$ test, $R^2$ and the adjusted $R^2$, the standard errors of the $B$s, the $t$ tests, and significance levels for the $B$s.

As long as residuals are correlated only among themselves and *not* back with any of the predictor variables, the OLS regression coefficient estimates themselves should be unbiased. However, the $B$s are no longer *best* linear unbiased estimators, and the estimates of the statistical significance of the $B$s and the constant term are inaccurate. If there is positive autocorrelation (the more usual case in trend data), the $t$ tests will be inappropriately large. If there is negative autocorrelation (less common), the computer program's calculated $t$ tests will be too small. However, there also may be autocorrelation in an independent variable, which generally aggravates the underestimation of residuals in OLS regression.

*Diagnosing First-Order Autocorrelation*

There are several ways to detect first-order autocorrelation in least squares analyses. Pairs of adjacent residuals can be plotted against time (or space) and the resulting scatterplot examined. However, the scatterplot "cloud of points" mentioned in most introductory statistics texts often resembles just that, especially with large samples. The decision is literally based on an "eyeball" analysis.

Second, and more formally, the statistical significance of the number of positive and negative runs or sign changes in the residuals can be tested. Tables of significance tests for the runs test are available in many statistics textbooks. The situation of too many runs means the adjacent residuals have switched signs too often and oscillate, resulting in a diagnosis of negative autocorrelation. The situation of too few runs means long streams of positive or negative trends, thus suggesting positive autocorrelation. The number of runs expected in a random progression of elements depends on the number of observations. Most tables apply to relatively small sample sizes, such as $N < 40$. Since many time series for social trends are relatively short in duration, depending on the availability of data, this test can be more practical than it initially appears.

One widely used formal diagnostic for first-order autocorrelation is the Durbin-Watson $d$ statistic, which is available in many statistical computer programs. The $d$ statistic is approximately calculated as $2(1-\rho)$ where $\rho e_t e_{t-1}$ is the intraclass correlation coefficient. The $e_t$ can be defined as adjacent residuals (in the following formula, $v$ represents the true random error terms that one really wants to estimate):

$$e_{\mathrm{t}} = \rho e_{t-1} + v_t$$

Thus $d$ is a ratio of the sum of squared differences between adjacent residuals to the sum of squared residuals. The $d$ has an interesting statistical distribution: Values near 2 imply $\rho = 0$ (no autocorrelation); $d$ is 0 when $\rho = 1$ (extreme positive autocorrelation) and 4 when $\rho = -1$ (extreme negative autocorrelation). In addition, $d$ has two *zones of indecision* (one near 0 and one near 4), in which the null hypothesis $\rho = 0$ is neither accepted nor rejected. The zones of indecision depend on the number of cases and the number of predictor variables. The $d$ calculation cannot be used with regressions through the origin, with standardized regression equations, or with equations that include lags of the dependent variable as predictors.

Many other computer programs provide iterative estimates of $\rho$ and its standard error, and sometimes the Durbin–Watson $d$ as well. Hierarchical linear models and time-series analysis programs are two examples. The null hypothesis $\rho = 0$ can be tested through a $t$-distribution with the ratio

$$\rho / \mathrm{se}_\rho.$$

The $t$ value can be evaluated using the $t$ tables if needed. If $\rho$ is not statistically significant, there is no first-order autocorrelation. If the analyst is willing to specify the positive or negative direction of the autocorrelation in advance, one-tailed tests of statistical significance are available.

## Possible Solutions

When interest centers on a time series and the lag of the dependent variable, it is tempting to attempt solving the autocorrelation problem by simply including a lagged dependent variable (e.g., $y_{t-1}$) as a predictor in OLS regression or as a covariate in analysis of covariance. Unfortunately, this alternative creates a worse problem. Because the observations are correlated, the residual term $e$ is now correlated back with $y_{t-1}$, which is a predictor for the regression or analysis of covariance. Not only does this alternative introduce bias into the previously unbiased $B$ coefficient estimates, but using lags also invalidates the use of diagnostic tests such as the Durbin-Watson $d$.

The first-differences (Cochrane-Orcutt) solution is one way to correct autocorrelation. This generalized least squares (GLS) solution creates a set of new variables by subtracting from each variable (not just the dependent variable) its own $t-1$ lag or adjacent case. Then each newly created variable in the equation is multiplied by the weight $(1-\rho)$ to make the error terms behave randomly.

An analyst may also wish to check for higher order autoregressive processes. If a GLS solution was created for the AR (1) autocorrelation, some

statistical programs will test for the statistical significance of $\rho$ using the Durbin-Watson $d$ for the reestimated GLS equation. If $\rho$ does not equal 0, higher order autocorrelation may exist. Possible solutions here include logarithmic or polynomial transformations of the variables, which may attenuate $\rho$. The analyst may also wish to examine econometrics programs that estimate higher order autoregressive equations.

In the Cochrane-Orcutt solution, the first observation is lost; this may be problematic in small samples. The Prais-Winsten approximation has been used to estimate the first observation in case of bivariate correlation or regression (with a loss of one additional degree of freedom).

In most social and behavioral science data, once autocorrelation is corrected, conclusions about the statistical significance of the results become much more conservative. Even when corrections for $\rho$ have been made, some statisticians believe that $R^2$s or $\eta^2$s to estimate the total explained variance in regression or analysis of variance models are invalid if autocorrelation existed in the original analyses. The explained variance tends to be quite large under these circumstances, reflecting the covariation of trends or behaviors.

Several disciplines have other ways of handling autocorrelation. Some alternate solutions are paired $t$ tests and multivariate analysis of variance for either repeated measures or multiple dependent variables. Econometric analysts diagnose treatments of higher order periodicity, lags for either predictors or dependent variables, and moving averages (often called ARIMA). Specialized computer programs exist, either freestanding or within larger packages, such as the Statistical Package for the Social Sciences (SPSS; an IBM company, formerly called PASW® Statistics).

Autocorrelation is an unexpectedly common phenomenon that occurs in many social and behavioral science phenomena (e.g., psychological experiments or the tracking of student development over time, social trends on employment, or cluster samples). Its major possible consequence—leading one to believe that accidental sample fluctuations are statistically significant—is serious. Checking and correcting for autocorrelation should become a more automatic process in the data analyst's tool chest than it currently appears to be.

*Susan Carol Losh*

***See also*** Cluster Sampling; Hierarchical Linear Modeling; Intraclass Correlation; Multivariate Analysis of Variance (MANOVA); Time-Series Study

**Further Readings**

Bowerman, B. L. (2004). *Forecasting, time series, and regression* (4th ed.). Pacific Grove, CA: Duxbury Press.

Gujarati, D. (2009). *Basic econometrics* (5th ed.). New York: McGraw-Hill.

Luke, D. A. (2004). *Multilevel modeling.* Thousand Oaks, CA: Sage.

Menard, S. W. (2002). *Longitudinal research* (2nd ed.). Thousand Oaks, CA: Sage.

Ostrom, C. W., Jr. (1990), *Time series analysis: Regression techniques* (2nd ed.). Thousand Oaks, CA: Sage.

# B

## Bar Chart

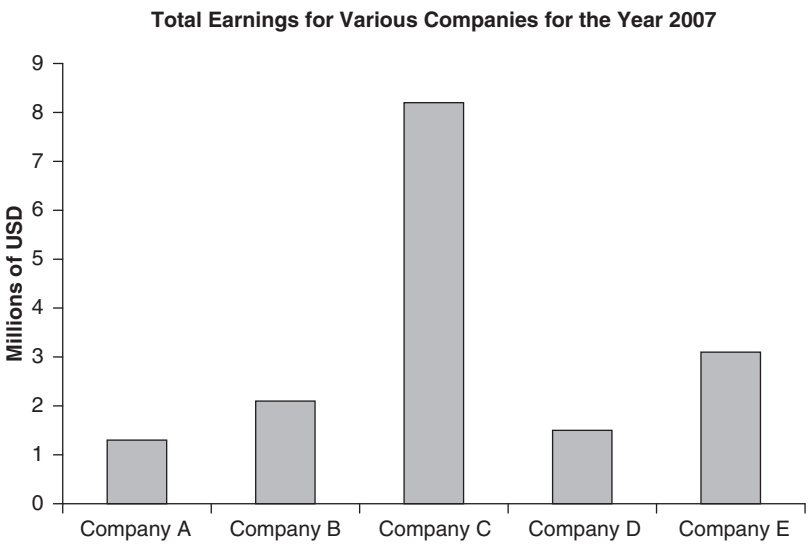
The term *bar chart* refers to a category of diagrams in which values are represented by the height or length of bars, lines, or other symbolic representations. Bar charts are typically used to display variables on a nominal or ordinal scale. Bar charts are a very popular form of information graphics often used in research articles, scientific reports, textbooks, and popular media to visually display relationships and trends in data. However, for this display to be effective, the data must be presented accurately, and the reader must be able to analyze the presentation effectively. This entry provides information on the history of the bar charts, the types of bar charts, and the construction of a bar chart.

### History

The creation of the first bar chart is attributed to William Playfair and appeared in *The Commercial and Political Atlas* in 1786. Playfair's bar graph was an adaptation of Joseph Priestley's time-line charts, which were popular at the time. Ironically, Playfair attributed his creation of the bar graph to a lack of data. In his *Atlas,* Playfair presented 34 plates containing line graphs or surface charts graphically representing the imports and exports from different countries over the years. Since he lacked the necessary time-series data for Scotland, he was forced to graph its trade data for a single year as a series of 34 bars, one for each of the imports and exports of Scotland's 17 trading partners. However, his innovation was largely ignored in Britain for a number of years. Playfair himself attributed little value to his invention, apologizing for what he saw as the limitations of the bar chart. It was not until 1801 and the publication of his *Statistical Breviary* that Playfair recognized the value of his invention. Playfair's invention fared better in Germany and France. In 1811 the German Alexander von Humboldt published adaptations of Playfair's bar graph and pie charts in *Essai Politique sur le Royaume de la Nouvelle Espagne*. In 1821, Jean Baptiste Joseph Fourier adapted the bar chart to create the first graph of cumulative frequency distribution, referred to as an ogive. In 1833, A. M. Guerry used the bar chart to plot crime data, creating the first histogram. Finally, in 1859 Playfair's work began to be accepted in Britain when Stanley Jevons published bar charts in his version of an economic atlas modeled on Playfair's earlier work. Jevons in turn influenced Karl Pearson, commonly considered the "father of modern statistics," who promoted the widespread acceptance of the bar chart and other forms of information graphics.

### Types

Although the terms *bar chart* and *bar graph* are now used interchangeably, the term *bar chart* was reserved traditionally for corresponding displays that did not have scales, grid lines, or tick marks.




| | |
|---|---|
| Company A | 1.3 |
| Company B | 2.1 |
| Company C | 8.2 |
| Company D | 1.5 |
| Company E | 3.1 |

**Figure 1** Simple Bar Chart and Associated Data

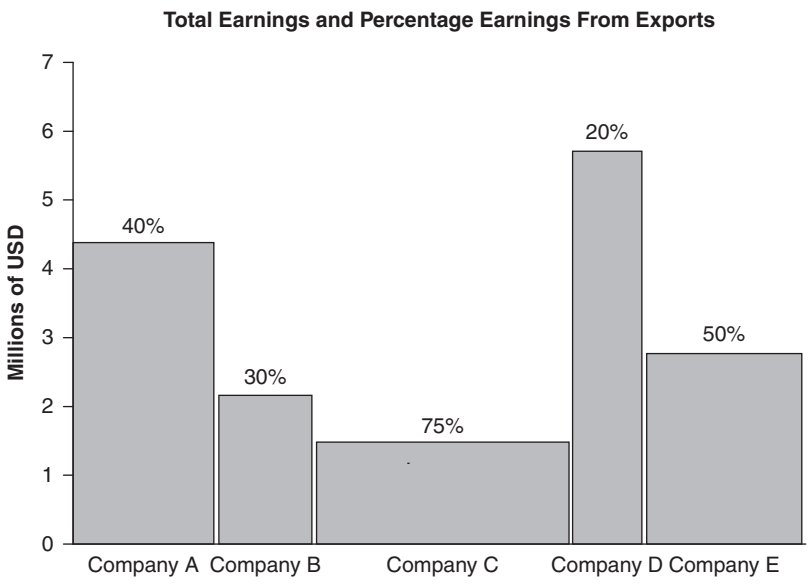
*Note:* USD = U.S. dollars.

The value each bar represented was instead shown on or adjacent to the data graphic.

An example bar chart is presented in Figure 1. Bar charts can display data by the use of either horizontal or vertical bars; vertical bar charts are also referred to as *column graphs*. The bars are typically of a uniform width with a uniform space between bars. The end of the bar represents the value of the category being plotted. When there is no space between the bars, the graph is referred to as a *joined bar graph* and is used to emphasize the differences between conditions or discrete categories. When continuous quantitative scales are used on both axes of a joined bar chart, the chart is referred to as a *histogram* and is often used to display the distribution of variables that are of interval or ratio scale. If the widths of the bars are not uniform but are instead used to display some measure or characteristic of the data element represented by the bar, the graph is referred to as an *area bar graph* (see Figure 2). In this graph, the heights of the bars represent the total earnings in U.S. dollars, and the widths of the bars are used to represent the percentage of the earnings coming from exports. The information expressed by the bar width can be displayed by means of a scale on the horizontal axis or by a legend, or, as in this case, the values might be noted directly on the graph. If both positive and negative values are plotted on the quantitative axis, the graph is called a *deviation graph*. On occasion the bars are replaced with pictures or symbols to make the graph more attractive or to visually represent the data series; these graphs are referred to as *pictographs* or *pictorial bar graphs*.

It may on occasion be desirable to display a confidence interval for the values plotted on the graph. In these cases the confidence intervals can be displayed by appending an error bar, or a shaded, striped, or tapered area to the end of the bar, representing the possible values covered in the confidence interval. If the bars are used to represent the range between the upper and lower values of a data series rather than one specific value, the graph is called a *range bar graph*. Typically the lower values are plotted on the left in a horizontal bar chart and on the bottom for a vertical bar chart. A line drawn across the bar can designate additional or inner values, such as a mean or median value. When the five-number summary (the minimum and maximum values, the upper and lower quartiles, and the median) is displayed, the graph is commonly referred to as a *box plot* or a *box-and-whisker diagram*.

A *simple bar graph* allows the display of a single data series, whereas a *grouped* or *clustered bar graph* displays two or more data series on one graph (see Figure 3). In clustered bar graphs, elements of the same category are plotted side by side; different colors, shades, or patterns, explained in a legend, may be used to differentiate the various data series, and the spaces between clusters distinguish the various categories.

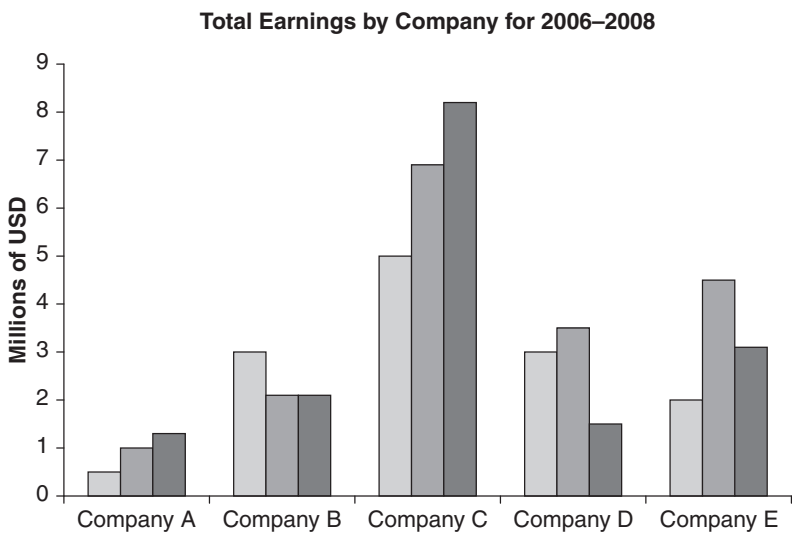



| | *Earnings (USD)* | *Percentage of Earnings from Exports* |
|---|---|---|
| Company A | 4.2 | 40 |
| Company B | 2.1 | 30 |
| Company C | 1.5 | 75 |
| Company D | 5.7 | 20 |
| Company E | 2.9 | 50 |

**Figure 2** Area Bar Graph and Associated Data

*Note:* USD = U.S. dollars.

While there is no limit to the number of series that can be plotted on the same graph, it is wise to limit the number of series plotted to no more than four in order to keep the graph from becoming confusing. To reduce the size of the graph and to improve readability, the bars for separate categories can be overlapped, but the overlap should be less than 75% to prevent the graph from being mistaken for a stacked bar graph. A *stacked bar graph,* also called a *divided* or *composite bar graph,* has multiple series stacked end to end instead of side by side. This graph displays the relative contribution of the components of a category; a different color, shade, or pattern differentiates each component, as described in a legend. The end of the bar represents the value of the whole category, and the heights of the various data series represent the relative contribution of the components of the category. If the graph represents the separate components' percentage of the whole value rather than the actual values, this graph is commonly referred to as a *100% stacked bar graph*. Lines can be drawn to connect the components of a stacked bar graph to more clearly delineate the relationship between the same components of different categories. A stacked bar graph can also use only one bar to demonstrate the contribution of the components of only one category, condition, or occasion, in which case it functions more like a pie chart. Two data series can also be plotted together in a *paired bar graph,* also referred to as a *sliding bar* or *bilateral bar graph*. This graph differs from a clustered bar graph because rather than being plotted side by side, the values for one data series are plotted with horizontal bars to the left and the values for the other data series are plotted with horizontal bars to the right. The units of measurement and scale intervals for the two data series need not be the same, allowing for a visual display of correlations and other meaningful relationships between the two data series. A paired bar graph can be a variation of either a simple, clustered, or stacked bar graph. A paired bar graph without spaces between the bars is often called a *pyramid graph* or a *two-way histogram*. Another method for comparing two data series is the *difference bar graph*. In this type of bar graph, the bars represent the difference in the values of two data series. For instance, one could compare the performance of two different classes on a series of tests or compare the different performance of males and females on a series of assessments. The direction of the difference can be noted at the ends of bars or by labeling the bars. When comparing multiple factors at two points in time or under two different conditions, one can use a *change bar graph*. The bars in this graph are used to represent the change between the two conditions or times. Since the direction of




| | *2006* | *2007* | *2008* |
|---|---|---|---|
| Company A | 0.5 | 1 | 1.3 |
| Company B | 3 | 2.1 | 2.1 |
| Company C | 5 | 6.9 | 8.2 |
| Company D | 3 | 3.5 | 1.5 |
| Company E | 2 | 4.5 | 3.1 |

**Figure 3** Clustered Bar Chart and Associated Data

*Note:* USD = U.S. dollars.

change is usually important with these types of graphs, a coding system is used to indicate the direction of the change.

## Creating an Effective Bar Chart

A well-designed bar chart can effectively communicate a substantial amount of information relatively easily, but a poorly designed graph can create confusion and lead to inaccurate conclusions among readers. Choosing the correct graphing format or technique is the first step in creating an effective graphical presentation of data. Bar charts are best used for making discrete comparisons between several categorical variables because the eye can spot very small differences in relative height. However, a bar chart works best with four to six categories; attempting to display more than six categories on a bar graph can lead to a crowded and confusing graph. Once an appropriate graphing technique has been chosen, it is important to choose the direction and the measurement scale for the primary axes. The decision to present the data in a horizontal or vertical format is largely a matter of personal preference; a vertical presentation, however, is more intuitive for displaying amount or quantity, and a horizontal presentation makes more sense for displaying distance or time. A horizontal presentation also allows for more space for detailed labeling of the categorical axis. The choice of an appropriate scale is critical for accurate presentation of data in a bar graph. Simple changes in the starting point or the interval of a scale can make the graph look dramatically different and may possibly misrepresent the relationships within the data. The best method for avoiding this problem is to always begin the quantitative scale at 0 and to use a linear rather than a logarithmic scale. However, in cases in which the values to be represented are extremely large, a start value of 0 effectively hides any differences in the data because by necessity the intervals must be extremely wide. In these cases it is possible to maintain smaller intervals while still starting the scale at 0 by the use of a clearly marked scale break. Alternatively, one can highlight the true relationship between the data by starting the scale at 0 and adding an inset of a small section of the larger graph to demonstrate the true relationship. Finally, it is important to make sure the graph and its axes are clearly labeled so that the reader can understand what data are being presented. Modern technology allows the addition of many superfluous graphical elements to enhance the basic graph design. Although the addition of these elements is a matter of personal choice, it is important to remember that the primary aim of data graphics is to display data accurately and clearly. If the additional elements detract from this clarity of presentation, they should be avoided.

*Teresa P. Clark and Sara E. Bolt*

*See also* Box-and-Whisker Plot; Distribution; Graphical Display of Data; Histogram; Pie Chart

**Further Readings**

Cleveland, W. S., & McGill, R. (1985). Graphical perception and graphical methods for analyzing scientific data. *Science, 229,* 828–833.
Harris, R. L. (1999). *Information graphics: A comprehensive illustrated reference.* New York: Oxford University Press.
Playfair, W. (1786). *The commercial and political atlas.* London: Corry.
Shah, P., & Hoeffner, J. (2002). Review of graph comprehension research: Implications for instruction. *Educational Psychology Review, 14,* 47–69.
Spence, I. (2000). The invention and use of statistical charts. *Journal de la Société Française de Statistique, 141,* 77–81.
Tufte, E. R. (1983). *The visual display of quantitative information.* Cheshire, CT: Graphics.
Wainer, H. (1996). Depicting error. *American Statistician, 50,* 101–111.

# Bartlett's Test

The assumption of equal variances across treatment groups may cause serious problems if violated in one-way analysis of variance models. A common test for homogeneity of variances is Bartlett's test. This statistical test checks whether the variances from different groups (or samples) are equal.

Suppose that there are $r$ treatment groups and we want to test

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_r^2$$

versus

$$H_1 : \sigma_m^2 \neq \sigma_k^2 \text{ for some } m \neq k.$$

In this context, we assume that we have independently chosen random samples of size $n_i, i = 1, \ldots, r$ from each of the $r$ independent populations. Let $X_{ij} \sim N(\mu_i, \sigma_i^2)$ be independently distributed with a normal distribution having mean $\mu_i$ and variance $\sigma_i^2$ for each $j = 1, \ldots, n_i$ and each $i = 1, \ldots, r$. Let $\overline{X}_i$ be the sample mean and $S_i^2$ the sample variance of the sample taken from the $i$th group or population. The uniformly most powerful unbiased parametric test of size $\alpha$ for testing for equality of variances among $r$ populations is known as Bartlett's test, and Bartlett's test statistic is given by

$$\ell_1^* = \frac{\prod_{i=1}^{r} (S_i^2)^{w_i}}{\sum_{i=1}^{r} w_i S_i^2},$$

where $w_i = (n_i - 1)/(N - r)$ is known as the weight for the $i$th group and $N = \sum_{i=1}^{r} n_i$ is the sum of the individual sample sizes. In the equireplicate case (i.e., $n_1 = \cdots = n_r = n$), the weights are equal, and $w_i = 1/r$ for each $i = 1, \ldots, r$. The test statistic is the ratio of the *weighted geometric mean* of the group sample variances to their *weighted arithmetic mean.* The values of the test statistic are bounded as $0 \leq \ell_1^* \leq 1$ by Jensen's inequality. Large values of $0 \leq \ell_1^* \leq 1$ (i.e., values near 1) indicate agreement with the null hypothesis, whereas small values indicate disagreement with the null. The terminology $\ell_1^*$ is used to indicate that Bartlett's test is based on M. S. Bartlett's modification of the likelihood ratio test, wherein he replaced the sample sizes $n_i$ with their corresponding degrees of freedom, $n_i - 1$. Bartlett did so to make the test unbiased. In the equireplicate case, Bartlett's test and the likelihood ratio test result in the same test statistic and same critical region.

The distribution of $\ell_1^*$ is complex even when the null hypothesis is true. R. E. Glaser showed that the distribution of $\ell_1^*$ could be expressed as a product of independently distributed beta random variables. In doing so he renewed much interest in the exact distribution of Bartlett's test. We reject $H_0$ provided $\ell_1^* \leq b_\alpha(n_1, \ldots, n_r)$ where $\Pr(\ell_1^* < b_\alpha(n_1, \ldots, n_r)) = \alpha$ when $H_0$ is true. The Bartlett critical value $b_\alpha(n_1, \ldots, n_r)$ is indexed by level of significance and the individual sample sizes. The critical values were first tabled in the equireplicate case, and the critical value was simplified to $b_\alpha(n, \ldots, n) = b_\alpha(n)$. Tabulating critical values with unequal sample sizes becomes counterproductive because of possible combinations of groups, sample sizes, and levels of significance.

## Example

Consider an experiment in which lead levels are measured at five different sites. The data in Table 1 come from Paul Berthouex and Linfield Brown:

**Table 1** Ten Measurements of Lead Concentration ($\mu G/L$) Measured on Waste Water Specimens

| | *Measurement No.* | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| *Lab* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 | 3.4 | 3.0 | 3.4 | 5.0 | 5.1 | 5.5 | 5.4 | 4.2 | 3.8 | 4.2 |
| 2 | 4.5 | 3.7 | 3.8 | 3.9 | 4.3 | 3.9 | 4.1 | 4.0 | 3.0 | 4.5 |
| 3 | 5.3 | 4.7 | 3.6 | 5.0 | 3.6 | 4.5 | 4.6 | 5.3 | 3.9 | 4.1 |
| 4 | 3.2 | 3.4 | 3.1 | 3.0 | 3.9 | 2.0 | 1.9 | 2.7 | 3.8 | 4.2 |
| 5 | 3.3 | 2.4 | 2.7 | 3.2 | 3.3 | 2.9 | 4.4 | 3.4 | 4.8 | 3.0 |

*Source:* Berthouex, P. M., & Brown, L. C. (2002). *Statistics for environmental engineers* (2nd ed., p. 170). Boca Raton, FL: Lewis.

From these data one can compute the sample variances and weights, which are given below:

| *Labs* | *Weight* | *Variance* |
|---|---|---|
| 1 | 0.2 | 0.81778 |
| 2 | 0.2 | 0.19344 |
| 3 | 0.2 | 0.41156 |
| 4 | 0.2 | 0.58400 |
| 5 | 0.2 | 0.54267 |

By substituting these values into the formula for $\ell_1^*$, we obtain

$$\ell_1^* = \frac{0.46016}{0.509889} = 0.90248.$$

Critical values, $b_\alpha(\alpha, n)$, of Bartlett's test are tabled for cases in which the sample sizes are equal and $\alpha = .05$. These values are given in D. Dyer and Jerome Keating for various values of $r$, the number of groups, and $n$, the common sample size (see Table 2). Works by Glaser, M. T. Chao, and Glaser, and S. B. Nandi provide tables of exact critical values of Bartlett's test. The most extensive set is contained in Dyer and Keating. Extensions (for larger numbers of groups) to the table of critical values can be found in Keating, Glaser, and N. S. Ketchum.

## Approximation

In the event that the sample sizes are not equal, one can use the *Dyer–Keating approximation* to the critical values:

$$b_\alpha(a; n_1, \ldots, n_a) \doteq \sum_{i=1}^{a} \frac{n_i}{N} \times b_\alpha(a, n_i).$$

So for the lead levels, we have the following values: $b_{0.05}(5; 10) \doteq 0.8025$. At the 5% level of significance, there is not enough evidence to reject the null hypothesis of equal variances.

Because of the complexity of the distribution of $\ell_1^*$, Bartlett's test originally employed an approximation. Bartlett proved that

$$\left[-\ln\left(\ell_1^*\right)\right]/c \sim \chi^2(r-1),$$

where

$$c = \frac{1 + \left[\frac{1}{3(r-1)}\right] \sum_{i=1}^{r} \frac{1}{n_i - 1} - \frac{1}{N-r}}{N-r}.$$

The approximation works poorly for small sample sizes. This approximation is more accurate as sample sizes increase, and it is recommended that $\min(n_i) \geq 3$ and that most $n_i > 5$.

## Assumptions

Bartlett's test statistic is quite sensitive to nonnormality. In fact, W. J. Conover, M. E. Johnson, and M. M. Johnson echo the results of G. E. P. Box that Bartlett's test is very sensitive to samples that exhibit nonnormal kurtosis. They recommend that Bartlett's test be used only when the data conform to normality. Prior to using Bartlett's test, it is recommended that one test for normality using an appropriate test such as the Shapiro–Wilk W test. In the event that the normality assumption is violated, it is recommended that one test equality of variances using Howard Levene's test.

*Mark T. Leung and Jerome P. Keating*

***See also*** Critical Value; Likelihood Ratio Statistic; Normality Assumption; Parametric Statistics; Variance

**Table 2** Table of Bartlett's Critical Values

| | *Number of Populations*, r | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *n* | *2* | *3* | *4* | *5* | *6* | *7* | *8* | *9* | *10* |
| *3* | .3123 | .3058 | .3173 | .3299 | . | . | . | . | . |
| *4* | .4780 | .4699 | .4803 | .4921 | .5028 | .5122 | .5204 | .5277 | .5341 |
| *5* | .5845 | .5762 | .5850 | .5952 | .6045 | .6126 | .6197 | .6260 | .6315 |
| *6* | .6563 | .6483 | .6559 | .6646 | .6727 | .6798 | .6860 | .6914 | .6961 |
| *7* | .7075 | .7000 | .7065 | .7142 | .7213 | .7275 | .7329 | .7376 | .7418 |
| *8* | .7456 | .7387 | .7444 | .7512 | .7574 | .7629 | .7677 | .7719 | .7757 |
| *9* | .7751 | .7686 | .7737 | .7798 | .7854 | .7903 | .7946 | .7984 | .8017 |
| *10* | .7984 | .7924 | .7970 | .8025 | .8076 | .8121 | .8160 | .8194 | .8224 |
| *11* | .8175 | .8118 | .8160 | .8210 | .8257 | .8298 | .8333 | .8365 | .8392 |
| *12* | .8332 | .8280 | .8317 | .8364 | .8407 | .8444 | .8477 | .8506 | .8531 |
| *13* | .8465 | .8415 | .8450 | .8493 | .8533 | .8568 | .8598 | .8625 | .8648 |
| *14* | .8578 | .8532 | .8564 | .8604 | .8641 | .8673 | .8701 | .8726 | .8748 |
| *15* | .8676 | .8632 | .8662 | .8699 | .8734 | .8764 | .8790 | .8814 | .8834 |
| *16* | .8761 | .8719 | .8747 | .8782 | .8815 | .8843 | .8868 | .8890 | .8909 |
| *17* | .8836 | .8796 | .8823 | .8856 | .8886 | .8913 | .8936 | .8957 | .8975 |
| *18* | .8902 | .8865 | .8890 | .8921 | .8949 | .8975 | .8997 | .9016 | .9033 |
| *19* | .8961 | .8926 | .8949 | .8979 | .9006 | .9030 | .9051 | .9069 | .9086 |
| *20* | .9015 | .8980 | .9003 | .9031 | .9057 | .9080 | .9100 | .9117 | .9132 |
| *21* | .9063 | .9030 | .9051 | .9078 | .9103 | .9124 | .9143 | .9160 | .9175 |
| *22* | .9106 | .9075 | .9095 | .9120 | .9144 | .9165 | .9183 | .9199 | .9213 |
| *23* | .9146 | .9116 | .9135 | .9159 | .9182 | .9202 | .9219 | .9235 | .9248 |
| *24* | .9182 | .9153 | .9172 | .9195 | .9217 | .9236 | .9253 | .9267 | .9280 |
| *25* | .9216 | .9187 | .9205 | .9228 | .9249 | .9267 | .9283 | .9297 | .9309 |
| *26* | .9246 | .9219 | .9236 | .9258 | .9278 | .9296 | .9311 | .9325 | .9336 |
| *27* | .9275 | .9249 | .9265 | .9286 | .9305 | .9322 | .9337 | .9350 | .9361 |
| *28* | .9301 | .9276 | .9292 | .9312 | .9330 | .9347 | .9361 | .9374 | .9385 |
| *29* | .9326 | .9301 | .9316 | .9336 | .9354 | .9370 | .9383 | .9396 | .9406 |
| *30* | .9348 | .9325 | .9340 | .9358 | .9376 | .9391 | .9404 | .9416 | .9426 |
| *40* | .9513 | .9495 | .9506 | .9520 | .9533 | .9545 | .9555 | .9564 | .9572 |
| *50* | .9612 | .9597 | .9606 | .9617 | .9628 | .9637 | .9645 | .9652 | .9658 |
| *60* | .9677 | .9665 | .9672 | .9681 | .9690 | .9698 | .9705 | .9710 | .9716 |
| *80* | .9758 | .9749 | .9754 | .9761 | .9768 | .9774 | .9779 | .9783 | .9787 |
| *100* | .9807 | .9799 | .9804 | .9809 | .9815 | .9819 | .9823 | .9827 | .9830 |

*Source:* Dyer, D., & Keating, J. P. (1980). On the determination of critical values for Bartlett's test. *Journal of the American Statistical Association, 75,* 313–319. Reprinted with permission from the *Journal of the American Statistical Association.* Copyright 1980 by the American Statistical Association. All rights reserved.

*Note:* The table shows the critical values used in Bartlett's test of equal variance at the 5% level of significance.

**Further Readings**

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Statistical Society, Series A, 160,* 268–282.

Berthouex, P. M., & Brown, L. C. (2002). *Statistics for environmental engineers* (2nd ed.). Boca Raton, FL: Lewis.

Box, G. E. P. (1953). Nonnormality and tests on variances. *Biometrika, 40,* 318–335.

Chao, M. T., & Glaser, R. E. (1978). The exact distribution of Bartlett's test statistic for homogeneity of variances with unequal sample sizes. *Journal of the American Statistical Association, 73,* 422–426.

Conover, W. J., Johnson, M. E., & Johnson, M. M. (1981). A comparative study of tests of homogeneity of variances with applications to the Outer Continental Shelf bidding data. *Technometrics, 23,* 351–361.

Dyer, D., & Keating, J. P. (1980). On the determination of critical values for Bartlett's test. *Journal of the American Statistical Association, 75,* 313–319.

Glaser, R. E. (1976). Exact critical values for Bartlett's test for homogeneity of variances. *Journal of the American Statistical Association, 71,* 488–490.

Keating, J. P., Glaser, R. E., & Ketchum, N. S. (1990). Testing hypotheses about the shape parameter of a gamma distribution. *Technometrics, 32,* 67–82.

Levene, H. (1960). Robust tests for equality of variances. In I. Olkin (Ed.), *Contributions to probability and statistics* (pp. 278–292). Palo Alto, CA: Stanford University Press.

Madansky, A. (1989). *Prescriptions for working statisticians*. New York: Springer-Verlag.

Nandi, S. B. (1980). On the exact distribution of a normalized ratio of weighted geometric mean to the unweighted arithmetic mean in samples from gamma distributions. *Journal of the American Statistical Association, 75,* 217–220.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika, 52,* 591–611.

# Barycentric Discriminant Analysis

Barycentric discriminant analysis (BADIA) generalizes discriminant analysis, and like discriminant analysis, it is performed when measurements made on some observations are combined to assign these observations, or *new* observations, to a priori defined categories. For example, BADIA can be used (a) to assign people to a given diagnostic group (e.g., patients with Alzheimer's disease, patients with other dementia, or people aging without dementia) on the basis of brain imaging data or psychological tests (here the a priori categories are the clinical groups), (b) to assign wines to a region of production on the basis of several physical and chemical measurements (here the a priori categories are the regions of production), (c) to use brain scans taken on a given participant to determine what type of object (e.g., a face, a cat, a chair) was watched by the participant when the scans were taken (here the a priori categories are the types of object), or (d) to use DNA measurements to predict whether a person is at risk for a given health problem (here the a priori categories are the types of health problem).

BADIA is more general than standard discriminant analysis because it can be used in cases for which discriminant analysis cannot be used. This is the case, for example, when there are more variables than observations or when the measurements are categorical.

BADIA is a class of methods that all rely on the same principle: Each category of interest is represented by the *barycenter* of its observations (i.e., the weighted average; the barycenter is also called the *center of gravity* of the observations of a given category), and a generalized principal components analysis (GPCA) is performed on the category by variable matrix. This analysis gives a set of discriminant factor scores for the categories and another set of factor scores for the variables. The original observations are then projected onto the category factor space, providing a set of factor scores for the observations. The distance of each observation to the set of categories is computed from the factor scores, and each observation is assigned to the closest category. The comparison between the a priori and a posteriori category assignments is used to assess the quality of the discriminant procedure. The prediction for the observations that were used to compute the barycenters is called the *fixed-effect* prediction. Fixed-effect performance is evaluated by counting the number of correct and incorrect assignments and storing these numbers in a confusion matrix. Another index of the performance of the fixed-effect model—equivalent to a squared coefficient

of correlation—is the ratio of category variance to the sum of category variance plus variance of the observations within each category. This coefficient is denoted $R^2$ and is interpreted as the proportion of variance of the observations explained by the categories or as the proportion of the variance explained by the discriminant model. The performance of the fixed-effect model can also be represented graphically as a *tolerance* ellipsoid that encompasses a given proportion (say 95%) of the observations. The overlap between the tolerance ellipsoids of two categories is proportional to the number of misclassifications between these two categories.

New observations can also be projected onto the discriminant factor space, and they can be assigned to the closest category. When the actual assignment of these observations is not known, the model can be used to *predict* category membership. The model is then called a *random* model (as opposed to the fixed model). An obvious problem, then, is to evaluate the quality of the prediction for new observations. Ideally, the performance of the random-effect model is evaluated by counting the number of correct and incorrect classifications for new observations and computing a confusion matrix on these new observations. However, it is not always practical or even feasible to obtain new observations, and therefore the random-effect performance is, in general, evaluated using computational cross-validation techniques such as the *jackknife* or the *bootstrap*. For example, a jackknife approach (also called *leave one out*) can be used by which each observation is taken out of the set, in turn, and predicted from the model built on the other observations. The predicted observations are then projected in the space of the fixed-effect discriminant scores. This can also be represented graphically as a *prediction* ellipsoid. A prediction ellipsoid encompasses a given proportion (say 95%) of the new observations. The overlap between the prediction ellipsoids of two categories is proportional to the number of misclassifications of new observations between these two categories.

The stability of the discriminant model can be assessed by a cross-validation model such as the bootstrap. In this procedure, multiple sets of observations are generated by sampling with replacement from the original set of observations, and the category barycenters are computed from each of these sets. These barycenters are then projected onto the discriminant factor scores. The variability of the barycenters can be represented graphically as a *confidence* ellipsoid that encompasses a given proportion (say 95%) of the barycenters. When the confidence intervals of two categories do not overlap, these two categories are *significantly* different.

In summary, BADIA is a GPCA performed on the category barycenters. GPCA encompasses various techniques, such as correspondence analysis, biplot, Hellinger distance analysis, discriminant analysis, and canonical variate analysis. For each specific type of GPCA, there is a corresponding version of BADIA. For example, when the GPCA is correspondence analysis, this is best handled with the most well-known version of BADIA: discriminant correspondence analysis. Because BADIA is based on GPCA, it can also analyze data tables obtained by the concatenation of blocks (i.e., subtables). In this case, the importance (often called the *contribution*) of each block to the overall discrimination can also be evaluated and represented as a graph.

*Hervé Abdi and Lynne J. Williams*

***See also*** Bootstrapping; Canonical Correlation Analysis; Correspondence Analysis; Discriminant Analysis; Jackknife; Matrix Algebra; Principal Components Analysis

**Further Readings**

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Greenacre, M. J. (1984). *Theory and applications of correspondence analysis*. London: Academic Press.

Saporta, G., & Niang, N. (2006). Correspondence analysis and classification. In M. Greenacre & J. Blasius (Eds.), *Multiple correspondence analysis and related methods* (pp. 371–392). Boca Raton, FL: Chapman & Hall/CRC.

# Bayes's Theorem

Bayes's theorem is a simple mathematical formula used for calculating conditional probabilities. It figures prominently in subjectivist or Bayesian

approaches to statistics, epistemology, and inductive logic. Subjectivists, who maintain that rational belief is governed by the laws of probability, lean heavily on conditional probabilities in their theories of evidence and their models of empirical learning. Bayes's theorem is central to these paradigms because it simplifies the calculation of conditional probabilities and clarifies significant features of the subjectivist position.

This entry begins with a brief history of Thomas Bayes and the publication of his theorem. Next, the entry focuses on probability and its role in Bayes's theorem. Last, the entry explores modern applications of Bayes's theorem.

## History

Thomas Bayes was born in 1702, probably in London, England. Others have suggested the place of his birth to be Hertfordshire. He was the eldest of six children of Joshua and Ann Carpenter Bayes. His father was a nonconformist minister, one of the first seven in England. Information on Bayes's childhood is scarce. Some sources state that he was privately educated, and others state he received a liberal education to prepare for the ministry. After assisting his father for many years, he spent his adult life as a Presbyterian minister at the chapel in Tunbridge Wells. In 1742, Bayes was elected as a fellow by the Royal Society of London. He retired in 1752 and remained in Tunbridge Wells until his death in April of 1761.

Throughout his life he wrote very little, and only two of his works are known to have been published. These two essays are *Divine Benevolence,* published in 1731, and *Introduction to the Doctrine of Fluxions,* published in 1736. He was known as a mathematician not for these essays but for two other papers he had written but never published. His studies focused in the areas of probability and statistics. His posthumously published article now known by the title "An Essay Towards Solving a Problem in the Doctrine of Chances" developed the idea of inverse probability, which later became associated with his name as Bayes's theorem. Inverse probability was so called because it involves inferring backwards from the data to the parameter (i.e., from the effect to the cause). Initially, Bayes's ideas attracted little attention. It was not until after the French mathematician Pierre-Simon Laplace published his paper "Mémoire sur la Probabilité des Causes par les Évènements" in 1774 that Bayes's ideas gained wider attention. Laplace extended the use of inverse probability to a variety of distributions and introduced the notion of "indifference" as a means of specifying prior distributions in the absence of prior knowledge. Inverse probability became during the 19th century the most commonly used method for making statistical inferences. Some of the more famous examples of the use of inverse probability to draw inferences during this period include estimation of the mass of Saturn, the probability of the birth of a boy at different locations, the utility of antiseptics, and the accuracy of judicial decisions.

In the latter half of the 19th century, authorities such as Siméon-Denis Poisson, Bernard Bolzano, Robert Leslie Ellis, Jakob Friedrich Fries, John Stuart Mill, and A. A. Cournot began to make distinctions between probabilities about things and probabilities involving our beliefs about things. Some of these authors attached the terms *objective* and *subjective* to the two types of probability. Toward the end of the century, Karl Pearson, in his *Grammar of Science,* argued for using experience to determine prior distributions, an approach that eventually evolved into what is now known as *empirical Bayes*. The Bayesian idea of inverse probability was also being challenged toward the end of the 19th century, with the criticism focusing on the use of uniform or "indifference" prior distributions to express a lack of prior knowledge.

The criticism of Bayesian ideas spurred research into statistical methods that did not rely on prior knowledge and the choice of prior distributions. In 1922, Ronald Alymer Fisher's paper "On the Mathematical Foundations of Theoretical Statistics," which introduced the idea of likelihood and maximum likelihood estimates, revolutionized modern statistical thinking. Jerzy Neyman and Egon Pearson extended Fisher's work by adding the ideas of hypothesis testing and confidence intervals. Eventually the collective work of Fisher, Neyman, and Pearson became known as *frequentist* methods. From the 1920s to the 1950s, frequentist methods displaced inverse probability as the primary methods used by researchers to make statistical inferences.

Interest in using Bayesian methods for statistical inference revived in the 1950s, inspired by Leonard Jimmie Savage's 1954 book *The Foundations of Statistics*. Savage's work built on previous work of several earlier authors exploring the idea of subjective probability, in particular the work of Bruno de Finetti. It was during this time that the terms *Bayesian* and *frequentist* began to be used to refer to the two statistical inference camps. The number of papers and authors using Bayesian statistics continued to grow in the 1960s. Examples of Bayesian research from this period include an investigation by Frederick Mosteller and David Wallace into the authorship of several of the Federalist papers and the use of Bayesian methods to estimate the parameters of time-series models. The introduction of Monte Carlo Markov chain (MCMC) methods to the Bayesian world in the late 1980s made computations that were impractical or impossible earlier realistic and relatively easy. The result has been a resurgence of interest in the use of Bayesian methods to draw statistical inferences.

## Publishing of Bayes's Theorem

Bayes never published his mathematical papers, and therein lies a mystery. Some suggest his theological concerns with modesty might have played a role in his decision. However, after Bayes's death, his family asked Richard Price to examine Bayes's work. Price was responsible for the communication of Bayes's essay on probability and chance to the Royal Society. Although Price was making Bayes's work known, he was occasionally mistaken for the author of the essays and for a time received credit for them. In fact, Price only added introductions and appendixes to works he had published for Bayes, although he would eventually write a follow-up paper to Bayes's work.

The present form of Bayes's theorem was actually derived not by Bayes but by Laplace. Laplace used the information provided by Bayes to construct the theorem in 1774. Only in later papers did Laplace acknowledge Bayes's work.

## Inspiration of Bayes's Theorem

In "An Essay Towards Solving a Problem in the Doctrine of Chances," Bayes posed a problem to be solved: "*Given* the number of times in which an unknown event has happened and failed: *Required* the chance that the probability of its happening in a single trial lies somewhere between any two degrees of probability that can be named." Bayes's reasoning began with the idea of conditional probability:

If $P(B) > 0$, the conditional probability of $A$ given $B$, denoted by $P(A|B)$, is

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \text{ or } \frac{P(AB)}{P(B)}.$$

Bayes's main focus then became defining $P(B|A)$ in terms of $P(A|B)$.

A key component that Bayes needed was the *law of total probability*. Sometimes it is not possible to calculate the probability of the occurrence of an event $A$. However, it is possible to find $P(A|B)$ and $P(A|B^c)$ for some event $B$where $B^c$ is the complement of $B$. The weighted average, $P(A)$, of the probability of $A$ given that $B$ has occurred and the probability of $A$ given that $B$ has not occurred can be defined as follows:

Let $B$ be an event with $P(B) > 0$ and $P(B^c) > 0$. Then for any event $A$,

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c).$$

If there are $k$ events, $B1, \ldots, B_k$, that form a partition of the sample space, and $A$ is another event in the sample space, then the events $B_1A, B_2A, \ldots, B_kA$ will form a partition of $A$. Thus, the law of total probability can be extended as follows:

Let $B_j$ be an event with $P(Bj) > 0$ for $j = 1, \ldots, k$. Then for any event $A$,

$$P(A) = \sum_{j=1}^{k} P(B_j)P(A|B_j).$$

These basic rules of probability served as the inspiration for Bayes's theorem.

## Bayes's Theorem

Bayes's theorem allows for a reduction in uncertainty by considering events that have occurred. The theorem is applicable as long as the probability of the more recent event (given an earlier event) is known. With this theorem, one can find the

probability of the earlier event, given the more recent event that has occurred. The earlier event is known as the *prior* probability. The primary focus is on the probability of the earlier event given the more recent event that has occurred (called the *posterior* probability). The theorem can be described in the following manner:

Let $B_j$ be an event with $P(B_j) > 0$ for $j = 1, \ldots, k$ and forming a partition of the sample space. Furthermore, let $A$ be an event such that $P(A) > 0$. Then for $i = 1, \ldots, k$,

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{\sum_{j=1}^{k} P(B_j)P(A|B_j)}.$$

$P(B_i)$ is the *prior* probability and the probability of the earlier event. $P(A|B_i)$ is the probability of the more recent event given the prior has occurred and is referred to as the *likelihood*. $P(B_i|A)$ is what one is solving for and is the probability of the earlier event given that the recent event has occurred (the posterior probability). There is also a version of Bayes's theorem based on a secondary event $C$ :

$$P(B_i|AC) = \frac{P(B_i|C)P(A|B_iC)}{\sum_{j=1}^{k} P(B_j|C)P(A|B_jC)}.$$

***Example***

A box contains 7 red and 13 blue balls. Two balls are selected at random and are discarded without their colors being seen. If a third ball is drawn randomly and observed to be red, what is the probability that both of the discarded balls were blue?

Let $BB$, $BR$, and $RR$ represent the events that the discarded balls are blue and blue, blue and red, and red and red, respectively. Let $R$ represent the event that the third ball chosen is red. Solve for the posterior probability $P(BB|R)$.

$$P(BB|R) = \frac{P(R|BB)P(BB)}{P(R|BB)P(BB) + P(R|BR)P(BR) + P(R|RR)P(RR)}$$

The probability that the first two balls drawn were blue, red, or blue and red are 39/95, 21/190, and 91/190, in that order. Now,

$$P(BB|R) = \frac{\frac{7}{18} * \frac{39}{95}}{\frac{7}{18} * \frac{39}{95} + \frac{6}{18} * \frac{91}{190} + \frac{5}{18} * \frac{21}{190}} \approx 0.46$$

The probability that the first two balls chosen were blue given the third ball selected was red is approximately 46%.

Bayes's theorem can be applied to the real world to make appropriate estimations of probability in a given situation. Diagnostic testing is one example in which the theorem is a useful tool. Diagnostic testing identifies whether a person has a particular disease. However, these tests contain error. Thus, a person can test positive for the disease and in actuality not be carrying the disease. Bayes's theorem can be used to estimate the probability that a person truly has the disease given that the person tests positive. As an illustration of this, suppose that a particular cancer is found for every 1 person in 2,000. Furthermore, if a person has the disease, there is a 90% chance the diagnostic procedure will result in a positive identification. If a person does not have the disease, the test will give a false positive 1% of the time. Using Bayes's theorem, the probability that a person with a positive test result, actually has the cancer C, is

$$P(C|P) = \frac{\frac{1}{2000}(.90)}{\frac{1}{2000}(0.90) + \frac{1999}{2000}(0.01)} \approx 0.043.$$

If a person tests positive for the cancer test, there is only a 4% chance that the person has the cancer. Consequently, follow-up tests are almost always necessary to verify a positive finding with medical screening tests.

Bayes's theorem has also been used in psychometrics to make a classification scale rather than an ability scale in the classroom. A simple example of classification is dividing a population into two categories of mastery and nonmastery of a subject. A test would be devised to determine whether a person falls in the mastery or the nonmastery category. The posterior probabilities for different skills can be collected, and the results would show mastered skills and nonmastered skills that need attention. The test may even allow for new posterior probabilities to be computed after each question.

The two examples presented above are just a small sample of the applications in which Bayes's theorem has been useful. While certain academic fields concentrate on its use more than others do, the theorem has far-reaching influence in business, medicine, education, psychology, and so on.

## Bayesian Statistical Inference

Bayes's theorem provides a foundation for Bayesian statistical inference. However, the approach to inference is different from that of a traditional (frequentist) point of view. With Bayes's theorem, inference is dynamic. That is, a Bayesian approach uses evidence about a phenomenon to update knowledge of prior beliefs.

There are two popular ways to approach inference. The traditional way is the frequentist approach, in which the probability $P$ of an uncertain event $A$, written $P(A)$, is defined by the frequency of that event, based on previous observations. In general, population parameters are considered as fixed effects and do not have distributional form. The frequentist approach to defining the probability of an uncertain event is sufficient, provided that one has been able to record accurate information about many past instances of the event. However, if no such historical database exists, then a different approach must be considered.

Bayesian inference is an approach that allows one to reason about beliefs under conditions of uncertainty. Different people may have different beliefs about the probability of a prior event, depending on their specific knowledge of factors that might affect its likelihood. Thus, Bayesian inference has no one correct probability or approach. Bayesian inference is dependent on both prior and observed data.

In a traditional hypothesis test, there are two complementary hypotheses: $H_0$, the status quo hypothesis, and $H_1$, the hypothesis of change. Letting $D$ stand for the observed data, Bayes's theorem applied to the hypotheses becomes

$$P(H_0|D) = \frac{P(H_0)P(D|H_0)}{P(H_0)P(D|H_0) + P(H_1)P(D|H_1)}$$

and

$$P(H_1|D) = \frac{P(H_1)P(D|H_1)}{P(H_0)P(D|H_0) + P(H_1)P(D|H_1)}.$$

The $P(H_0|D)$ and $P(H_1|D)$ are posterior probabilities (i.e., the probability that the null is true given the data and the probability that the alternative is true given the data, respectively). The $P(H_0)$ and $P(H_1)$ are prior probabilities (i.e., the probability that the null or the alternative is true prior to considering the new data, respectively).

In frequentist hypothesis testing, one considers only $P(D|H_0)$, which is called the *p value*. If the $p$ value is smaller than a predetermined significance level, then one rejects the null hypothesis and asserts the alternative hypothesis. One common mistake is to interpret the $p$ value as the probability that the null hypothesis is true, given the observed data. This interpretation is a Bayesian one. From a Bayesian perspective, one may obtain $P(H_0|D)$, and if that probability is sufficiently small, then one rejects the null hypothesis in favor of the alternative hypothesis. In addition, with a Bayesian approach, several alternative hypotheses can be considered at one time.

As with traditional frequentist confidence intervals, a *credible interval* can be computed in Bayesian statistics. This credible interval is defined as the posterior probability interval and is used in ways similar to the uses of confidence intervals in frequentist statistics. For example, a 95% credible interval means that the posterior probability of the parameter lying in the given range is 0.95. A frequentist 95% confidence interval means that with a large number of repeated samples, 95% of the calculated confidence intervals would include the true value of the parameter; yet the probability that the parameter is inside the actual calculated confidence interval is either 0 or 1. In general, Bayesian credible intervals do not match a frequentist confidence interval, since the credible interval incorporates information from the prior distribution whereas confidence intervals are based only on the data.

## Modern Applications

In Bayesian statistics, information about the data and a priori information are combined to

estimate the posterior distribution of the parameters. This posterior distribution is used to infer the values of the parameters, along with the associated uncertainty. Multiple tests and predictions can be performed simultaneously and flexibly. Quantities of interest that are functions of the parameters are straightforward to estimate, again including the uncertainty. Posterior inferences can be updated as more data are obtained, so study design is more flexible than for frequentist methods.

Bayesian inference is possible in a number of contexts in which frequentist methods are deficient. For instance, Bayesian inference can be performed with small data sets. More broadly, Bayesian statistics is useful when the data set may be large but when few data points are associated with a particular treatment. In such situations standard frequentist estimators can be inappropriate because the likelihood may not be well approximated by a normal distribution. The use of Bayesian statistics also allows for the incorporation of prior information and for simultaneous inference using data from multiple studies. Inference is also possible for complex hierarchical models.

Lately, computation for Bayesian models is most often done via MCMC techniques, which obtain dependent samples from the posterior distribution of the parameters. In MCMC, a set of initial parameter values is chosen. These parameter values are then iteratively updated via a specially constructed Markovian transition. In the limit of the number of iterations, the parameter values are distributed according to the posterior distribution. In practice, after approximate convergence of the Markov chain, the time series of sets of parameter values can be stored and then used for inference via empirical averaging (i.e., Monte Carlo). The accuracy of this empirical averaging depends on the effective sample size of the stored parameter values, that is, the number of iterations of the chain after convergence, adjusted for the autocorrelation of the chain. One method of specifying the Markovian transition is via Metropolis–Hastings, which proposes a change in the parameters, often according to a random walk (the assumption that many unpredictable small fluctuations will occur in a chain of events), and then accepts or rejects that move with a probability that is dependent on the current and proposed state.

In order to perform valid inference, the Markov chain must have approximately converged to the posterior distribution before the samples are stored and used for inference. In addition, enough samples must be stored after convergence to have a large effective sample size; if the autocorrelation of the chain is high, then the number of samples needs to be large. Lack of convergence or high autocorrelation of the chain is detected via convergence diagnostics, which include autocorrelation and trace plots, as well as Geweke, Gelman–Rubin, and Heidelberger–Welch diagnostics. Software for MCMC can also be validated by a distinct set of techniques. These techniques compare the posterior samples drawn by the software with samples from the prior and the data model, thereby validating the joint distribution of the data and parameters as estimated by the software.

*Brandon K. Vaughn and Daniel L. Murphy*

***See also*** Estimation; Hypothesis; Inference: Deductive and Inductive; Parametric Statistics; Probability, Laws of

**Further Readings**

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London, 53,* 370–418.

Box, G. E., & Jenkins, G. M. (1970). *Time series analysis: Forecasting and control.* San Francisco: Holden-Day.

Dale, A. I. (1991). *A history of inverse probability from Thomas Bayes to Karl Pearson.* New York: Springer-Verlag.

Daston, L. (1994). How probabilities came to be objective and subjective. *Historia Mathematica, 21,* 330–344.

Fienberg, S. E. (1992). A brief history of statistics in three and one-half chapters: A review essay. *Statistical Science, 7,* 208–225.

Fienberg, S. E. (2006). When did Bayesian analysis become "Bayesian"? *Bayesian Analysis, 1*(1), 1–40.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, Series A, 222,* 309–368.

Laplace, P. S. (1774). Mémoire sur la probabilité des causes par les évènements [Memoir on the probability

of causes of events]. *Mémoires de la Mathématique et de Physique Presentés à l'Académie Royale des Sciences, Par Divers Savans, & Lûs dans ses Assemblées, 6,* 621–656.
Mosteller, F., & Wallace, D. L. (1963). Inference in an authorship problem. *Journal of the American Statistical Association, 59,* 275–309.
Pearson, K. (1892). *The grammar of science.* London: Walter Scott.
Savage, L. J. (1954). *The foundations of statistics.* New York: Wiley.
Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900.* Cambridge, MA: Harvard University Press.

## Behavior Analysis Design

Behavior analysis is a specific scientific approach to studying behavior that evolved from John Watson's behaviorism and the operant research model popularized by B. F. Skinner during the middle of the 20th century. This approach stresses direct experimentation and measurement of observable behavior. A basic assumption of behavior analysis is that behavior is malleable and controlled primarily by consequences. B. F. Skinner described the basic unit of behavior as an *operant,* a behavior emitted to operate on the environment. Additionally, he proposed that the response rate of the operant serve as the basic datum of the scientific study of behavior. An operant is characterized by a response that occurs within a specific environment and produces a specific consequence. According to the principles of operant conditioning, behavior is a function of three interactive components, illustrated by the three-term contingency: context, response, and consequences of behavior. The relationship between these three variables forms the basis of all behavioral research. Within this framework, individual components of the three-term contingency can be studied by manipulating experimental context, response requirements, or consequences of behavior. A change in any one of these components often changes the overall function of behavior, resulting in a change in future behavior. If the consequence strengthens future behavior, the process is called *reinforcement*. If future behavior is weakened or eliminated as a result of changing the consequence of behavior, the process is called *punishment.*

Behavior analysis encompasses two types of research: the experimental analysis of behavior, consisting of research to discover basic underlying behavioral principles, and applied behavior analysis, involving research implementing basic principles in real-world situations. Researchers in this field are often referred to as behavior analysts, and their research can take place in both laboratory and naturalistic settings and with animals and humans. Basic behavioral processes can be studied in any species, and the findings may be applied to other species. Therefore, researchers can use animals for experimentation, which can increase experimental control by eliminating or reducing confounding variables. Since it is important to verify that findings generalize across species, experiments are often replicated with other animals and with humans. Applied behavior analysis strives to develop empirically based interventions rooted in principles discovered through basic research. Many empirically based treatments have been developed with participants ranging from children with autism to corporate executives and to students and substance abusers. Contributions have been made in developmental disabilities, mental retardation, rehabilitation, delinquency, mental health, counseling, education and teaching, business and industry, and substance abuse and addiction, with potential in many other areas of social significance. Similar designs are employed in both basic and applied research, but they differ with regard to subjects studied, experimental settings, and degree of environmental control.

Regardless of the subject matter, a primary feature of behavior analytic research is that the behavior of individual organisms is examined under conditions that are rigorously controlled. One subject can provide a representative sample, and studying an individual subject thoroughly can sometimes provide more information than can studying many subjects because each subject's data are considered an independent replication. Behavior analysts demonstrate the reliable manipulation of behavior by changing the environment. Manipulating the environment allows researchers to discover the relationships between behavior and environment. This method is referred to as *single-subject* or *within-subject research* and requires

unique designs, which have been outlined by James Johnston and Hank Pennypacker. Consequently, this method takes an approach to the collection, validity, analysis, and generality of data that is different from approaches that primarily use group designs and inferential statistics to study behavior.

## Measurement Considerations

### *Defining Response Classes*

Measurement in single-subject design is objective and restricted to observable phenomena. Measurement considerations can contribute to behavioral variability that can obscure experimental effects, so care must be taken to avoid potential confounding variables. Measurement focuses on targeting a *response class,* which is any set of responses that result in the same environmental change. Response classes are typically defined by function rather than topography. This means that the form of the responses may vary considerably but produce the same result. For example, a button can be pressed several ways, with one finger, with the palm, with the toe, or with several fingers. The exact method of action is unimportant, but any behavior resulting in button depression is part of a response class. Topographical definitions are likely to result in classes that include some or all of several functional response classes, which can produce unwanted variability. Researchers try to arrange the environment to minimize variability within a clearly defined response class.

There are many ways to quantify the occurrence of a response class member. The characteristics of the behavior captured in its definition must suit the needs of the experiment, be able to address the experimental question, and meet practical limits for observation. In animal studies, a response is typically defined as the closing of a circuit in an experimental chamber by depressing a lever or pushing a key or button. With this type of response, the frequency and duration that a circuit is closed can be recorded. Conditions can be arranged to measure the force used to push the button or lever, the amount of time that occurs between responses, and the latency and accuracy of responding in relation to some experimentally arranged stimulus. These measurements serve as dependent variables. In human studies, the response is typically more broadly defined and may be highly individualized. For example, self-injurious behavior in a child with autism may include many forms that meet a common definition of minimum force that leaves a mark. Just as in basic research, a variety of behavioral measurements can be used as dependent variables. The response class must be sensitive to the influence of the independent variable (IV) without being affected by extraneous variables so that effects can be detected. The response class must be defined in such a way that researchers can clearly observe and record behavior.

### *Observation and Recording*

Once researchers define a response class, the methods of observation and recording are important in order to obtain a complete and accurate record of the subject's behavior. Measurement is direct when the focus of the experiment is the same as the phenomenon being measured. Indirect measurement is typically avoided in behavioral research because it undermines experimental control. Mechanical, electrical, or electronic devices can be used to record responses, or human observers can be selected and trained for data collection. Machine and human observations may be used together throughout an experiment. Behavior is continuous, so observational procedures must be designed to detect and record each response within the targeted response class.

## Experimental Design and Demonstration of Experimental Effects

### *Experimental Arrangements*

The most basic single-subject experimental design is the baseline–treatment sequence, the AB design. This procedure cannot account for certain confounds, such as maturation, environmental history, or unknown extraneous variables. Replicating components of the AB design provide additional evidence that the IV is the source of any change in the dependent measure. Replication designs consist of a *baseline* or *control* condition (A), followed by one or more *experimental* or *treatment* conditions (B), with additional conditions indicated by successive letters. Subjects experience both the control and the experimental

conditions, often in sequence and perhaps more than once. An ABA design replicates the original baseline, while an ABAB design replicates the baseline and the experimental conditions, allowing researchers to infer causal relationships between variables. These designs can be compared with a light switch. The first time one moves the switch from the on position to the off position, one cannot be completely certain that one's behavior was responsible for the change in lighting conditions. One cannot be sure the light bulb did not burn out at that exact moment or the electricity did not shut off coincidentally. Confidence is bolstered when one pushes the switch back to the on position and the lights turn back on. With a replication of moving the switch to off again, one has total confidence that the switch is controlling the light.

Single-subject research determines the effectiveness of the IV by eliminating or holding constant any potential confounding sources of variability. One or more behavioral measures are used as dependent variables so that data comparisons are made from one condition to another. Any change in behavior between the control and the experimental conditions is attributed to the effects of the IV. The outcome provides a detailed interpretation of the effects of an IV on the behavior of the subject.

Replication designs work only in cases in which effects are reversible. Sequence effects can occur when experience in one experimental condition affects a subject's behavior in subsequent conditions. The researcher must be careful to ensure consistent experimental conditions over replications. Multiple-baseline designs with multiple individuals, multiple behaviors, or multiple settings can be used in circumstances in which sequence effects occur, or as a variation on the AB design. Results are compared across control and experimental conditions, and factors such as irreversibility of effects, maturation of the subject, and sequence effect can be examined.

### *Behavioral Variability*

Variability in single-subject design refers both to variations in features of responding within a single response class and to variations in summary measures of that class, which researchers may be examining across sessions or entire phases of the experiment. The causes of variability can often be identified and systematically evaluated. Behavior analysts have demonstrated that frequently changing the environment results in greater degrees of variability. Inversely, holding the environment constant for a time allows behavior to stabilize and minimizes variability. Murray Sidman has offered several suggestions for decreasing variability, including strengthening the variables that directly maintain the behavior of interest, such as increasing deprivation, increasing the intensity of the consequences, making stimuli more detectable, or providing feedback to the subject. If these changes do not immediately affect variability, it could be that behavior requires exposure to the condition for a longer duration. Employing these strategies to control variability increases the likelihood that results can be interpreted and replicated.

### *Reduction of Confounding Variables*

Extraneous, or confounding, variables affect the detection of behavioral change due to the IV. Only by eliminating or minimizing external sources of variability can data be judged as accurately reflecting performance. Subjects should be selected that are similar along extra-experimental dimensions in order to reduce extraneous sources of variability. For example, it is common practice to use animals from the same litter or to select human participants on the basis of age, level of education, or socioeconomic status. Environmental history of an organism can also influence the target behav ior; therefore, subject selection methods should attempt to minimize differences between subjects. Some types of confounding variables cannot be removed, and the researcher must design an experiment to minimize their effects.

### *Steady State Behavior*

Single-subject designs rely on the collection of steady state baseline data prior to the administration of the IV. Steady states are obtained by exposing the subject to only one condition consistently until behavior stabilizes over time. Stabilization is determined by graphically examining the variability in behavior. Stability can be defined as a pattern of responding that exhibits relatively little

variation in its measured dimensional quantities over time.

Stability criteria specify the standards for evaluating steady states. Dimensions of behavior such as duration, latency, rate, and intensity can be judged as stable or variable during the course of experimental study, with rate most commonly used to determine behavioral stability. Stability criteria must set limits on two types of variability over time. The first is systematic increases and decreases of behavior, or *trend,* and the second is unsystematic changes in behavior, or *bounce*. Only when behavior is stable, without trend or bounce, should the next condition be introduced. Specific stability criteria include time, visual inspection of graphical data, and simple statistics. Time criteria can designate the number of experimental sessions or discrete period in which behavior stabilizes. The time criterion chosen must encompass even the slowest subject. A time criterion allowing for longer exposure to the condition may needlessly lengthen the experiment if stability occurs rapidly; on the other hand, behavior might still be unstable, necessitating experience and good judgment when a time criterion is used. A comparison of steady state behavior under baseline and different experimental conditions allows researchers to examine the effects of the IV.

## Scientific Discovery Through Data Analysis

Single-subject designs use visual comparison of steady state responding between conditions as the primary method of data analysis. Visual analysis usually involves the assessment of several variables evident in graphed data. These variables include upward or downward trend, the amount of variability within and across conditions, and differences in means and stability both within and across conditions. Continuous data are displayed against the smallest unit of time that is likely to show systematic variability. Cumulative graphs provide the greatest level of detail by showing the distribution of individual responses over time and across various stimulus conditions. Data can be summarized with less precision by the use of descriptive statistics such as measures of central tendency (mean and median), variation (interquartile range and standard deviation), and association (correlation and linear regression). These methods obscure individual response variability but can highlight the effects of the experimental conditions on responding, thus promoting steady states. Responding summarized across individual sessions represents some combination of individual responses across a group of sessions, such as mean response rate during baseline conditions. This method should not be the only means of analysis but is useful when one is looking for differences among sets of sessions sharing common characteristics.

Single-subject design uses ongoing behavioral data to establish steady states and make decisions about the experimental conditions. Graphical analysis is completed throughout the experiment, so any problems with the design or measurement can be uncovered immediately and corrected. However, graphical analysis is not without criticism. Some have found that visual inspection can be insensitive to small but potentially important differences of graphic data. When evaluating the significance of data from this perspective, one must take into account the magnitude of the effect, variability in data, adequacy of experimental design, value of misses and false alarms, social significance, durability of behavior change, and number and kinds of subjects. The best approach to analysis of behavioral data probably uses some combination of both graphical and statistical methods because each approach has relative advantages and disadvantages.

### *Judging Significance*

Changes in level, trend, variability, and serial dependency must be detected in order for one to evaluate behavioral data. *Level* refers to the general magnitude of behavior for some specific dimension. For example, 40 responses per minute is a lower level than 100 responses per minute. *Trend* refers to the increasing or decreasing nature of behavior change. *Variability* refers to changes in behavior from measurement to measurement. *Serial dependency* occurs when a measurement obtained during one time period is related to a value obtained earlier.

Several features of graphs are important, such as trend lines, axis units, number of data points, and condition demarcation. Trend lines are lines that fit the data best within a condition. These lines allow for discrimination of level and may

assist in discrimination of behavioral trends. The axis serves as an anchor for data, and data points near the bottom of a graph are easier to interpret than data in the middle of a graph. The number of data points also seems to affect decisions, with fewer points per phase improving accuracy.

### Generality

Generality, or how the results of an individual experiment apply in a broader context outside the laboratory, is essential to advancing science. The dimensions of generality include subjects, response classes, settings, species, variables, methods, and processes. Single-subject designs typically involve a small number of subjects that are evaluated numerous times, permitting in-depth analysis of these individuals and the phenomenon in question, while providing *systematic replication*. Systematic replication enhances generality of findings to other populations or conditions and increases internal validity. The *internal validity* of an experiment is demonstrated when additional subjects demonstrate similar behavior under similar conditions; although the absolute level of behavior may vary among subjects, the relationship between the IV and the relative effect on behavior has been reliably demonstrated, illustrating generalization.

*Jennifer L. Bredthauer and Wendy D. Donlin-Washington*

***See also*** Animal Research; Applied Research; Experimental Design; Graphical Display of Data; Independent Variable; Research Design Principles; Single-Subject Design; Trend Analysis; Within-Subjects Design

#### Further Readings

Bailey, J. S., & Burch, M. R. (2002). *Research methods in applied behavior analysis*. Thousand Oaks, CA: Sage.

Baron, A., & Perone, M. (1998). Experimental design and analysis in the laboratory of human operant behavior. In K. A. Lattal & M. Perone (Eds.), *Handbook of methods in human operant behavior* (pp. 3–14). New York: Plenum Press.

Fisch, G. S. (1998). Visual inspection of data revisited: Do the eyes still have it? *Behavior Analyst, 21*(1), 111–123.

Johnston, J. M., & Pennypacker, H. S. (1993). *Strategies and tactics of behavioral research* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Mazur, J. E. (2007). *Learning and behavior* (6th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Poling, A., & Grossett, D. (1986). Basic research designs in applied behavior analysis. In A. Poling & R. W. Fuqua (Eds.), *Research methods in applied behavior analysis: Issues and advances* (pp. 7–27). New York: Plenum Press.

Sidman, M. (1960). *Tactics of scientific research: Evaluating experimental data in psychology*. Boston: Authors Cooperative.

Skinner, B. F. (1938). *The behavior of organisms: An experimental analysis*. New York: D. Appleton-Century.

Skinner, B. F. (1965). *Science and human behavior*. New York: Free Press.

Watson, J. B. (1913). Psychology as a behaviorist views it. *Psychological Review, 20,* 158–177.

## Behrens–Fisher $t'$ Statistic

The Behrens–Fisher $t'$ statistic can be employed when one seeks to make inferences about the means of two normal populations without assuming the variances are equal. The statistic was offered first by W. U. Behrens in 1929 and reformulated by Ronald A. Fisher in 1939:

$$t' = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} = t_1 \sin\theta - t_2 \cos\theta,$$

where sample mean $\overline{x}_1$ and sample variance $s_1^2$ are obtained from the random sample of size $n_1$ from the normal distribution with mean $\mu_1$ and variance $\sigma_1^2$, $t_1 = (\overline{x}_1 - \mu_1)/\sqrt{s_1^2/n_1}$ has a $t$ distribution with $\nu_1 = n_1 - 1$ degrees of freedom, the respective quantities with subscript 2 are defined similarly, and $\tan\theta = \left(s_1/\sqrt{n_1}\right)/\left(s_2/\sqrt{n_2}\right)$ or $\theta = \tan^{-1}\left[\left(s_1/\sqrt{n_1}\right)/\left(s_2/\sqrt{n_2}\right)\right]$. The distribution of $t'$ is the Behrens–Fisher distribution. It is, hence, a mixture of the two $t$ distributions. The problem arising when one tries to test the normal population means without making any assumptions about their variances is referred to as the Behrens–Fisher problem or as the two means problem.

Under the usual null hypothesis of $H_0 : \mu_1 = \mu_2$, the test statistic $t'$ can be obtained

and compared with the percentage points of the Behrens–Fisher distribution. Tables for the Behrens–Fisher distribution are available, and the table entries are prepared on the basis of the four numbers $\nu_1 = n_1 - 1$, $\nu_2 = n_2 - 1$, $\theta$, and the Type I error rate $\alpha$. For example, Ronald A. Fisher and Frank Yates in 1957 presented significance points of the Behrens–Fisher distribution in two tables, one for $\nu_1$ and $\nu_2 = 6, 8, 12, 24, \infty$; $\theta = 0°$, $15°, 30°, 45°, 60°, 75°, 90°$; and $\alpha = .05, .01$, and the other for $\nu_1$ that is greater than $\nu_2 = 1, 2, 3, 4, 5, 6, 7$; $\theta = 0°, 15°, 30°, 45°, 60°, 75°$, $90°$ and $\alpha = .10, .05, .02, .01$. Seock-Ho Kim and Allan S. Cohen in 1998 presented significance points of the Behrens–Fisher distribution for $\nu_1$ that is greater than $\nu_2 = 2, 4, 6, 8, 10, 12$; $\theta = 0°$, $15°$, $30°$, $45°$, $60°, 75°, 90°$; and $\alpha = .10$, $.05, .02, .01$, and also offered computer programs for obtaining tail areas and percentage values of the Behrens–Fisher distribution.

Using the Behrens–Fisher distribution, one can construct the $100(1 - \alpha)\%$ interval that contains $\mu_1 - \mu_2$with

$$\overline{x}_1 - \overline{x}_2 \pm t'_{\alpha/2}(\nu_1, \nu_2, \theta)\sqrt{s_1^2/n_1 + s_2^2/n_2},$$

where the probability that $t' > t'_{\alpha/2}(\nu_1, \nu_2, \theta)$ is $\alpha/2$ or, equivalently, $\Pr\left[t' > t'_{\alpha/2}(\nu_1, \nu_2, \theta)\right] = \alpha/2$.

This entry first illustrates the statistic with an example. Then related methods are presented, and the methods are compared.

## Example

Driving times from a person's house to work were measured for two different routes with $n_1 = 5$ and $n_2 = 11$. The ordered data from the first route are 6.5, 6.8, 7.1, 7.3, 10.2, yielding $\overline{x}_1 = 7.580$ and $s_1^2 = 2.237$, and the data from the second route are 5.8, 5.8, 5.9, 6.0, 6.0, 6.0, 6.3, 6.3, 6.4, 6.5, 6.5, yielding $\overline{x}_2 = 6.136$ and $s_2^2 = 0.073$. It is assumed that the two independent samples were drawn from two normal distributions having means $\mu_1$and $\mu_2$ and variances $\sigma_1^2$ and $\sigma_2^2$, respectively. A researcher wants to know whether the average driving times differed for the two routes.

The test statistic under the null hypothesis of equal population means is $t' = 2.143$ with $\nu_1 = 4$, $\nu_2 = 10$, and $\theta = 83.078$. From the computer program, $\Pr(t' > 2.143) = .049$, indicating the null hypothesis cannot be rejected at $\alpha = .05$ when the alternative hypothesis is nondirectional, $H_a : \mu_1 \neq \mu_2$, because $p = .098$. The corresponding 95% interval for the population mean difference is $[-0.421, 3.308]$.

## Related Methods

The Student's $t$ test for independent means can be used when the two population variances are assumed to be equal and $\sigma_1^2 = \sigma_2^2 = \sigma^2$:

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2/n_1 + s_p^2/n_2}},$$

where the pooled variance that provides the estimate of the common population variance $\sigma^2$ is defined as $s_p^2 = \left[(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2\right]/(n_1 + n_2 - 2)$. It has a $t$ distribution with $\nu = n_1 + n_2 - 2$ degrees of freedom. The example data yield the Student's $t = 3.220$, $\nu = 14$, the two-tailed $p = .006$, and the 95% confidence interval of [0.482, 2.405]. The null hypothesis of equal population means is rejected at the nominal $\alpha = .05$, and the confidence interval does not contain 0.

When the two variances cannot be assumed to be the same, one of the solutions is to use the Behrens–Fisher $t'$ statistic. There are several alternative solutions. One simple way to solve the two means problem, called the smaller degrees of freedom $t$ test, is to use the same $t'$ statistic that has a $t$ distribution with different degrees of freedom:

$$t' \sim t[\min(\nu_1, \nu_2)],$$

where the degrees of freedom is the smaller value of $\nu_1$ or $\nu_2$. Note that this method should be used only if no statistical software is available because it yields a conservative test result and a wider confidence interval. The example data yield $t' = 2.143$, $\nu = 4$, the two-tailed $p = .099$, and the 95% confidence interval of $[-0.427, 3.314]$. The null hypothesis of equal population means is not rejected at $\alpha = .05$, and the confidence interval contains 0.

B. L. Welch in 1938 presented an approximate $t$ test. It uses the same $t'$ statistic that has

a $t$ distribution with the approximate degrees of freedom $\nu'$ :

$$t' \sim t(\nu'),$$

where $\nu' = 1 \Big/ \left[ c^2 / \nu_1 + (1-c)^2 \Big/ \nu_2 \right]$ with $c = \left(s_1^2/n_1\right) / \left[\left(s_1^2/n_1\right) + \left(s_2^2/n_2\right)\right]$. The approximation is accurate when both sample sizes are 5 or larger. Although there are other solutions, Welch's approximate $t$ test might be the best practical solution to the Behrens–Fisher problem because of its availability from the popular statistical software, including SPSS (an IBM company, formerly called PASW® Statistics) and SAS. The example data yield $t' = 2.143$, $\nu' = 4.118$, the two-tailed $p = .097$, and the 95% confidence interval of $[-0.406, 3.293]$ The null hypothesis of equal population means is not rejected at $\alpha = .05$, and the confidence interval contains 0.

In addition to the previous method, the Welch–Aspin $t$ test employs an approximation of the distribution of $t'$ by the method of moments. The example data yield $t' = 2.143$, and the critical value under the Welch–Aspin $t$ test for the two-tailed test is 2.715 at $\alpha = .05$. The corresponding 95% confidence interval is $[-0.386, 3.273]$. Again, the null hypothesis of equal population means is not rejected at $\alpha = .05$, and the confidence interval contains 0.

## Comparison of Methods

The Behrens–Fisher $t'$ statistic and the Behrens–Fisher distribution are based on Fisher's fiducial approach. The approach is to find a fiducial probability distribution that is a probability distribution of a parameter from observed data. Consequently, the interval that involves $t'_{\alpha/2}(\nu_1, \nu_2, \theta)$ is referred to as the $100(1-\alpha)\%$ fiducial interval.

The Bayesian solution to the Behrens–Fisher problem was offered by Harold Jeffreys in 1940. When uninformative uniform priors are used for the population parameters, the Bayesian solution to the Behrens–Fisher problem is identical to that of Fisher's in 1939. The Bayesian highest posterior density interval that contains the population mean difference with the probability of $1-\alpha$ is identical to the $100(1-\alpha)\%$ fiducial interval.

There are many solutions to the Behrens–Fisher problem based on the frequentist approach of Jerzy Neyman and Egon S. Pearson's sampling theory. Among the methods, Welch's approximate $t$ test and the Welch–Aspin $t$ test are the most important ones from the frequentist perspective. The critical values and the confidence intervals from various methods under the frequentist approach are in general different from those of either the fiducial or the Bayesian approach. For the one-sided alternative hypothesis, however, it is interesting to note that the generalized extreme region to obtain the generalized $p$ developed by Kam-Wah Tsui and Samaradasa Weerahandi in 1989 is identical to the extreme area from the Behrens–Fisher $t'$ statistic.

The critical values for the two-sided alternative hypothesis at $\alpha = .05$ for the example data are 2.776 for the smaller degrees of freedom $t$ test, 2.767 for the Behrens–Fisher $t'$ test, 2.745 for Welch's approximate $t$ test, 2.715 for the Welch–Aspin $t$ test, and 2.145 for the Student's $t$ test. The respective 95% fiducial and confidence intervals are $[-0.427, 3.314]$ for the smaller degrees of freedom test, $[-0.421, 3.308]$ for the Behrens–Fisher $t'$ test, $[-0.406, 3.293]$ for Welch's approximate $t$ test, $[-0.386, 3.273]$ for the Welch–Aspin $t$ test, and [0.482, 2.405] for the Student's $t$ test. The smaller degrees of freedom $t$ test yielded the most conservative result with the largest critical value and the widest confidence interval. The Student's $t$ test yielded the smallest critical value and the shortest confidence interval. All other intervals lie between these two intervals. The differences between many solutions to the Behrens–Fisher problem might be less than their differences from the Student's $t$ test when sample sizes are greater than 10.

The popular statistical software programs SPSS and SAS produce results from Welch's approximate $t$ test and the Student's $t$ test, as well as the respective confidence intervals. It is essential to have a table that contains the percentage points of the Behrens–Fisher distribution or computer programs that can calculate the tail areas and percentage values in order to use the Behrens–Fisher $t'$ test or to obtain the fiducial interval. Note that Welch's approximate $t$ test may not be as effective as the Welch–Aspin $t$ test. Note also that the sequential testing of the population means on the basis of the result from either *Levene's test* of the equal population variances from SPSS or the folded $F$ test

from SAS is not recommended in general because of the complicated nature of control of the Type I error (rejecting a true null hypothesis) in the sequential testing.

*Seock-Ho Kim*

***See also*** Mean Comparisons; Student's *t* Test; *t* Test, Independent Samples

**Further Readings**

Behrens, W. U. (1929). Ein Beitrag zur Fehlerberechnung bei wenigen Beobachtungen [A contribution to error estimation with few observations]. *Landwirtschaftliche Jahrbücher, 68,* 807–837.

Fisher, R. A. (1939). The comparison of samples with possibly unequal variances. *Annals of Eugenics, 9,* 174–180.

Fisher, R. A., & Yates, F. (1957). *Statistical tables for biological, agricultural and medical research* (4th ed.). Edinburgh, UK: Oliver and Boyd.

Jeffreys, H. (1940). Note on the Behrens-Fisher formula. *Annals of Eugenics, 10,* 48–51.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995). *Continuous univariate distributions* (Vol. 2, 2nd ed.). New York: Wiley.

Kendall, M., & Stuart, A. (1979). *The advanced theory of statistics* (Vol. 2, 4th ed.). New York: Oxford University Press.

Kim, S.-H., & Cohen, A. S. (1998). On the Behrens-Fisher problem: A review. *Journal of Educational and Behavioral Statistics, 23,* 356–377.

Tsui, K.-H., & Weerahandi, S. (1989). Generalized *p*-values in significance testing of hypotheses in the presence of nuisance parameters. *Journal of the American Statistical Association, 84,* 602–607; Correction, *86,* 256.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika, 29,* 350–362.

# Bernoulli Distribution

The Bernoulli distribution is a discrete probability distribution for a random variable that takes only two possible values, 0 and 1. Examples of events that lead to such a random variable include coin tossing (head or tail), answers to a test item (correct or incorrect), outcomes of a medical treatment (recovered or not recovered), and so on. Although it is the simplest probability distribution, it provides a basis for other important probability distributions, such as the binomial distribution and the negative binomial distribution.

## Definition and Properties

An experiment of chance whose result has only two possibilities is called a *Bernoulli trial* (or *Bernoulli experiment*). Let $p$ denote the probability of success in a Bernoulli trial $(0 < p < 1)$. Then, a random variable $X$ that assigns value 1 for a success with probability $p$ and value 0 for a failure with probability $1 - p$ is called a *Bernoulli random variable,* and it follows the Bernoulli distribution with probability $p$, which is denoted by $X \sim \text{Ber}(p)$. The probability mass function of $\text{Ber}(p)$ is given by

$$P(X = x) = p^x(1-p)^{1-x},\ x = 0, 1.$$

The mean of $X$ is $p$, and the variance is $p(1-p)$. Figure 1 shows the probability mass function of Ber(.7). The horizontal axis represents values of $X$, and the vertical axis represents the corresponding probabilities. Thus, the height is .7 at $X = 1$, and .3 for $X = 0$. The mean of Ber(0.7) is 0.7, and the variance is .21.

Suppose that a Bernoulli trial with probability $p$ is independently repeated for $n$ times, and we obtain a random sample $X_1, X_2, \ldots, X_n$. Then, the number of successes $Y = X_1 + X_2 + \cdots + X_n$ follows the *binomial distribution* with probability

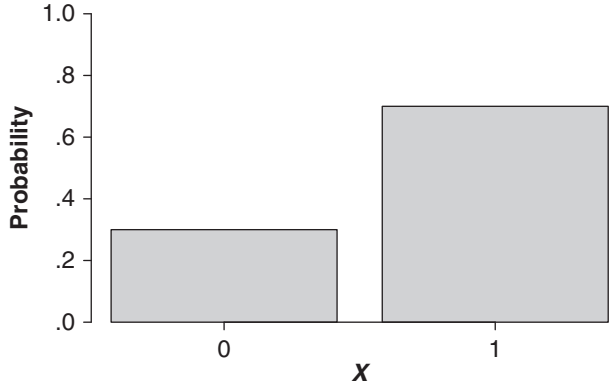



**Figure 1** Probability Mass Function of the Bernoulli Distribution With $p = .7$

$p$ and the number of trials $n$, which is denoted by $Y \sim \text{Bin}(n, p)$. Stated in the opposite way, the Bernoulli distribution is a special case of the binomial distribution in which the number of trials $n$ is 1. The probability mass function of Bin($n$,$p$) is given by

$$P(Y = y) = \frac{n!}{y!(n-y)!} p^y (1-p)^{n-y},$$
$$y = 0, 1, \ldots, n,$$

where $n!$ is the factorial of $n$, which equals the product $n(n-1)\cdots 2 \cdot 1$. The mean of $Y$ is $np$, and the variance is $np(1-p)$. Figure 2 shows the probability mass function of Bin(10,.7), which is obtained as the distribution of the sum of 10 independent random variables, each of which follows Ber(.7). The height of each bar represents the probability that $Y$ takes the corresponding value; for example, the probability of $Y = 7$ is about .27. The mean is 7 and the variance is 2.1. In general, the distribution is skewed to the right when $p < .5$, skewed to the left when $p > .5$, and symmetric when $p = .5$.

## Relationship to Other Probability Distributions

The Bernoulli distribution is a basis for many probability distributions, as well as for the binomial distribution. The number of failures before observing a success $t$ times in independent Bernoulli trials follows the *negative binomial distribution* with probability $p$ and the number of successes $t$. The *geometric distribution* is a special case of the negative binomial distribution in which the number of failures is counted before observing the first success (i.e., $t = 1$).

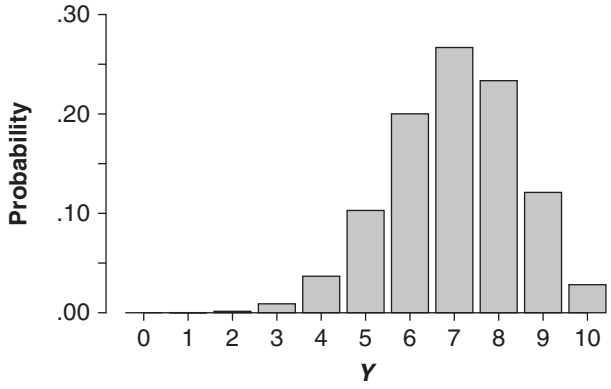



**Figure 2** Probability Mass Function of the Binomial Distribution With $p = 7$ and $n = 10$

Assume a finite Bernoulli population in which individual members are denoted by either 0 or 1. If sampling is done by randomly selecting one member at each time *with replacement* (i.e., each selected member is returned to the population before the next selection is made), then the resulting sequence constitutes independent Bernoulli trials, and the number of successes follows the binomial distribution. If sampling is done at random but *without* replacement, then each of the individual selections is still a Bernoulli trial, but they are no longer independent of each other. In this case, the number of successes follows the *hypergeometric distribution*, which is specified by the population probability $p$, the number of trials $n$, and the population size $m$.

Various approximations are available for the binomial distribution. These approximations are extremely useful when $n$ is large because in that case the factorials in the binomial probability mass function become prohibitively large and make probability calculations tedious. For example, by the central limit theorem, $Z = (Y - np)/\sqrt{np(1-p)}$ approximately follows the standard normal distribution $N(0, 1)$ when $Y \sim \text{Bin}(n, p)$. The constant 0.5 is often added to the denominator to improve the approximation (called *continuity correction*). As a rule of thumb, the normal approximation works well when either (a) $np(1-p) > 9$ or (b) $np > 9$ for $0 < p \leq .5$. The Poisson distribution with parameter $np$ also well approximates Bin($n$,$p$) when $n$ is large and $p$ is small. The Poisson approximation works well if $n^{0.31}p > .47$; for example, $p > .19, .14$, and .11 when $n = 20, 50$, and 100, respectively. If $n^{0.31}p \geq .47$, then the normal distribution gives better approximations.

## Estimation

Inferences regarding the population proportion $p$ can be made from a random sample $X_1, X_2, \ldots, X_n$ from Ber($p$), whose sum follows Bin($n$, $p$). The population proportion $p$ can be estimated by the sample mean (or the sample

proportion) $\hat{p} = \overline{X} = \sum_{i=1}^{n} X_i/n$, which is an unbiased estimator of $p$.

Interval estimation is usually made by the normal approximation. If $n$ is large enough (e.g., $n > 100$), a $100(1-\alpha)\%$ confidence interval is given by

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

where $\hat{p}$ is the sample proportion and $z_{\alpha/2}$ is the value of the standard normal variable that gives the probability $\alpha/2$ in the right tail. For smaller $n$s, the quadratic approximation gives better results:

$$\frac{1}{1+z_{\alpha/2}/n}\left(\hat{p} + \frac{z_{\alpha/2}^2}{2n} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z_{\alpha/2}^2}{4n^2}}\right).$$

The quadratic approximation works well if $.1 < p < .9$ and $n$ is as large as 25.

There are often cases in which one is interested in comparing two population proportions. Suppose that we obtained sample proportions $\hat{p}_1$ and $\hat{p}_2$ with sample sizes $n_1$ and $n_2$, respectively. Then, the difference between the population proportions is estimated by the difference between the sample proportions $\hat{p}_1 - \hat{p}_2$. Its standard error is given by

$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}},$$

from which one can construct a $100(1-\alpha)$ confidence interval as

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} SE(\hat{p}_1 - \hat{p}_2).$$

## Applications

### *Logistic Regression*

Logistic regression is a regression model about the Bernoulli probability and used when the dependent variable takes only two possible values. Logistic regression models are formulated as *generalized linear models* in which the canonical link function is the logit link and the Bernoulli distribution is assumed for the dependent variable.

In the standard case in which there are $K$ linear predictors $x_1, x_2, \ldots, x_K$ and the dependent variable $Y$, which represents a Bernoulli random variable (i.e., $Y = 0, 1$), the logistic regression model is expressed by the equation

$$\ln\frac{p(x)}{1-p(x)} = b_0 + b_1 x_1 + \cdots + b_K x_K,$$

where ln is the natural logarithm, $p(x)$ is the probability of $Y = 1$ (or the expected value of $Y$) given $x_1, x_2, \ldots, x_K$, and $b_0, b_1, \ldots, b_K$ are the regression coefficients. The left-hand side of the above equation is called the *logit*, or the *log-odds ratio*, of proportion $p$. The logit is symmetric about zero; it is positive (negative) if $p > .5$ ($p < .5$), and zero if $p = .5$. It approaches positive (negative) infinity as $p$ approaches 1 (0). Another representation equivalent to the above is

$$p(x) = \frac{\exp(b_0 + b_1 x_1 + \cdots + b_K x_K)}{1 + \exp(b_0 + b_1 x_1 + \cdots + b_K x_K)}.$$

The right-hand side is called the *logistic* regression function. In either case, the model states that the distribution of $Y$ given predictors $x_1, x_2, \ldots, x_K$ is Ber$[p(x)]$, where the logit of $p(x)$ is determined by a linear combination of predictors $x_1, x_2, \ldots, x_K$. The regression coefficients are estimated from $N$ sets of observed data $(Y_i, x_{i1}, x_{i2}, \ldots, x_{iK})$, $i = 1, 2, \ldots, N$.

### *The Binomial Error Model*

The binomial error model is one of the measurement models in the classical test theory. Suppose that there are $n$ test items, each of which is scored either 1 (correct) or 0 (incorrect). The binomial error model assumes that the distribution of person $i$'s total score $X_i$ given his or her "proportion-corrected" true score $\zeta_i (0 < \zeta_i < 1)$ is Bin$(n, \zeta_i)$:

$$P(X_i = x|\zeta_i) = \frac{n!}{x!(n-x)!}\zeta_i^x(1-\zeta_i)^{n-x},$$
$$x = 0, 1, \ldots, n.$$

This model builds on a simple assumption that for all items, the probability of a correct response for a person with true score $\zeta_i$ is equal to $\zeta_i$, but the error variance, $n\zeta_i(1-\zeta_i)$, varies as a function of $\zeta_i$ unlike the standard classical test model.

The observed total score $X_i = x_i$ serves as an estimate of $n\zeta_i$, and the associated error variance

can also be estimated as $\hat{\sigma}_i^2 = x_i(n - x_i)/(n - 1)$. Averaging this error variance over $N$ persons gives the overall error variance $\hat{\sigma}^2 = [\overline{x}(n - \overline{x}) - s^2]/(n - 1)$, where $\overline{x}$ is the sample mean of observed total scores over the $N$ persons and $s^2$ is the sample variance. It turns out that by substituting $\hat{\sigma}^2$ and $s^2$ in the definition of reliability, the reliability of the $n$-item test equals the Kuder–Richardson formula 21 under the binomial error model.

### History

The name *Bernoulli* was taken from Jakob Bernoulli, a Swiss mathematician in the 17th century. He made many contributions to mathematics, especially in calculus and probability theory. He is the first person who expressed the idea of the law of large numbers, along with its mathematical proof (thus, the law is also called *Bernoulli's theorem*). Bernoulli derived the binomial distribution in the case in which the probability $p$ is a rational number, and his result was published in 1713. Later in the 18th century, Thomas Bayes generalized Bernoulli's binomial distribution by removing its rational restriction on $p$ in his formulation of a statistical theory that is now known as Bayesian statistics.

*Kentaro Kato and William M. Bart*

***See also*** Logistic Regression; Normal Distribution; Odds Ratio; Poisson Distribution; Probability, Laws of

#### Further Readings

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York: Wiley.

Johnson, N. L., Kemp, A. W., & Kotz, S. (2005). *Univariate discrete distributions* (3rd ed.). Hoboken, NJ: Wiley.

Lindgren, B. W. (1993). *Statistical theory* (4th ed.). Boca Raton, FL: Chapman & Hall/CRC.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

## Beta

Beta ($\beta$) refers to the probability of Type II error in a statistical hypothesis test. Frequently, the power of a test, equal to $1 - \beta$ rather than $\beta$ itself, is referred to as a measure of quality for a hypothesis test. This entry discusses the role of $\beta$ in hypothesis testing and its relationship with significance ($\alpha$).

### Hypothesis Testing and Beta

Hypothesis testing is a very important part of statistical inference: the formal process of deciding whether a particular contention (called the *null hypothesis*) is supported by the data, or whether a second contention (called the *alternative hypothesis*) is preferred. In this context, one can represent the situation in a simple $2 \times 2$ decision table in which the columns reflect the true (unobservable) situation and the rows reflect the inference made based on a set of data:

| *Decision* | *Null Hypothesis Is True/Preferred* | *Alternative Hypothesis Is True/Preferred* |
|---|---|---|
| Fail to reject null hypothesis | Correct decision | Type II error |
| Reject null hypothesis in favor of alternative hypothesis | Type I error | Correct decision |

The language used in the decision table is subtle but deliberate. Although people commonly speak of accepting hypotheses, under the maxim that scientific theories are not so much proven as *supported* by evidence, we might more properly speak of failing to reject a hypothesis rather than of accepting it. Note also that it may be the case that neither the null nor the alternative hypothesis is, in fact, true, but generally we might think of one as preferable over the other on the basis of evidence. Semantics notwithstanding, the decision table makes clear that there exist two distinct possible types of error: that in which the null hypothesis is rejected when it is, in fact, true; and that in which the null hypothesis is not rejected when it is, in fact, false. A simple example that helps one in thinking about the difference between these two types of error is

a criminal trial in the U.S. judicial system. In that system, there is an initial presumption of innocence (null hypothesis), and evidence is presented in order to reach a decision to convict (reject the null hypothesis) or acquit (fail to reject the null). In this context, a Type I error is committed if an innocent person is convicted, while a Type II error is committed if a guilty person is acquitted. Clearly, both types of error cannot occur in a single trial; after all, a person cannot be both innocent and guilty of a particular crime. However, *a priori* we can conceive of the probability of each type of error, with the probability of a Type I error called the significance level of a test and denoted by $\alpha$, and the probability of a Type II error denoted by $\beta$, with $1 - \beta$, the probability of not committing a Type II error, called the *power of the test*.

## Relationship With Significance

Just as it is impossible to realize both types of error in a single test, it is also not possible to minimize both $\alpha$ and $\beta$ in a particular experiment with fixed sample size. In this sense, in a given experiment, there is a trade-off between $\alpha$ and $\beta$, meaning that both cannot be specified or guaranteed to be low. For example, a simple way to guarantee no chance of a Type I error would be to never reject the null hypothesis regardless of the data, but such a strategy would typically result in a very large $\beta$. Hence, it is common practice in statistical inference to fix the significance level at some nominal, low value (usually .05) and to compute and report $\beta$ in communicating the result of the test. Note the implied asymmetry between the two types of error possible from a hypothesis test: $\alpha$ is held at some prespecified value, while $\beta$ is not constrained. The preference for controlling $\alpha$ rather than $\beta$ also has an analogue in the judicial example above, in which the concept of "beyond reasonable doubt" captures the idea of setting $\alpha$ at some low level, and where there is an oft-stated preference for setting a guilty person free over convicting an innocent person, thereby preferring to commit a Type II error over a Type I error. The common choice of .05 for $\alpha$ most likely stems from Sir Ronald Fisher's 1926 statement that he "prefers to set a low standard of significance at the 5% point, and ignore entirely all results that fail to reach that level." He went on to say that "a scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance" (Fisher, 1926, p. 504).

Although it is not generally possible to control both $\alpha$ and $\beta$ for a test with a fixed sample size, it is typically possible to decrease $\beta$ while holding $\alpha$ constant if the sample size is increased. As a result, a simple way to conduct tests with high power (low $\beta$) is to select a sample size sufficiently large to guarantee a specified power for the test. Of course, such a sample size may be prohibitively large or even impossible, depending on the nature and cost of the experiment. From a research design perspective, sample size is the most critical aspect of ensuring that a test has sufficient power, and *a priori* sample size calculations designed to produce a specified power level are common when designing an experiment or survey. For example, if one wished to test the null hypothesis that a mean $\mu$ was equal to $\mu_0$ versus the alternative that $\mu$ was equal to $\mu_1 > \mu_0$, the sample size required to ensure a Type II error of $\beta$ if $\alpha = .05$ is $n = \{\sigma(1.645 - \mathbf{\Phi}^{-1}(\beta))/(\mu_1 - \mu_0)\}^2$, where $\mathbf{\Phi}$ is the standard normal cumulative distribution function and $\sigma$ is the underlying standard deviation, an estimate of which (usually the sample standard deviation) is used to compute the required sample size.

The value of $\beta$ for a test is also dependent on the effect size—that is, the measure of how different the null and alternative hypotheses are, or the size of the effect that the test is designed to detect. The larger the effect size, the lower $\beta$ will typically be at fixed sample size, or, in other words, the more easily the effect will be detected.

*Michael A. Martin and Steven Roberts*

***See also*** Hypothesis; Power; *p* Value; Type I Error; Type II Error

### Further Readings

Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain, 23,* 503–513.

Lehmann, E. L. (1986). *Testing statistical hypotheses.* New York: Wiley.

Moore, D. (1979). *Statistics: Concepts and controversies.* San Francisco: W. H. Freeman.

## Between-Subjects Design

*See* Single-Subject Design; Within-Subjects Design

## Bias

Bias is systematic error in data collected to address a research question. In contrast to random errors, which are randomly distributed and therefore even out across people or groups studied, biases are errors that are systematically related to people, groups, treatments, or experimental conditions and therefore cause the researcher to overestimate or underestimate the measurement of a behavior or trait. Bias is problematic because it can endanger the ability of researchers to draw valid conclusions about whether one variable causes a second variable (threats to internal validity) or whether the results generalize to other people (threats to external validity). Bias comes in many forms, including sampling bias, selection bias, experimenter expectancy effects, and response bias.

### Sampling Bias

Human participants in studies generally represent a subset of the entire population of people whom the researcher wishes to understand; this subset of the entire population is known as the study sample. Unless a study sample is chosen using some form of random sampling in which every member of the population has a known, nonzero chance of being chosen to participate in the study, it is likely that some form of sampling bias exists. Even for surveys that attempt to use random sampling of a population via random-digit dialing, the sample necessarily excludes people who do not have phone service. Thus people of lower socioeconomic status—and who are therefore less likely to be able to afford phone service—may be underrepresented in samples generated using random-digit dialing. (A modern artifact of random-digit dialing that leads to bias in the other direction is that those with cell phones and many people with landlines routinely screen calls and are less likely to answer the phone when the incoming call is from someone unknown.)

Researchers often rely on volunteers to participate in their studies, but there may be something different about those who volunteer to participate in studies and those who do not volunteer that systematically biases the sample. For example, people who volunteer to participate in studies of new treatments for psychological disorders may be more motivated to get better than those who do not volunteer, leading researchers to overestimate the effectiveness of a new treatment. Similarly, people who are selected to participate in surveys may choose not to respond to the survey. If there are systematic differences between responders and nonresponders, then the generalizability of the survey findings is limited.

### Selection Bias

Selection bias is present when participants in different study conditions possess different characteristics at the start of a study that could influence the outcome measures in the study. Selection bias is presumed in quasi-experimental designs, in which participants are not randomly assigned to experimental conditions, as is often the case in educational research in which classrooms or classes receive different interventions. For example, if a researcher wanted to examine whether students learn more in introductory psychology classes that have a small number of students than they do in classes with a large number, a selection bias may exist if better students are more likely to choose classes that have fewer students. If students in smaller classes outperform students in larger classes, it will be unclear whether the performance difference is the result of smaller classes or because better students self-select into smaller classes.

In this case, selection bias makes it appear that there are differences between the two conditions (large vs. small classes) when in truth there might be no difference (only differences between the types of people who sign up for the classes). Selection bias could also lead researchers to conclude that there are no differences between groups when in fact there are. Imagine a researcher wanted to test whether a new intervention reduced recidivism among juvenile offenders and that researcher relied on collaborators at social service agencies to

randomly assign participants to the intervention or the control group. If the collaborators sometimes broke with random assignment and assigned the juveniles who were most in need (e.g., had the worst criminal records) to the intervention group, then when both groups were subsequently followed to determine whether they continued to break the law (or were caught doing so), the selection bias would make it difficult to find a difference between the two groups. The preintervention differences in criminal behavior between the intervention and control groups might mask any effect of the intervention or even make it appear as if the intervention increased criminal behavior.

### Experimenter Expectancy Effects

Researchers usually have hypotheses about how subjects will perform under different experimental conditions. When a researcher knows which experimental group a subject is assigned to, the researcher may unintentionally behave differently toward the participant. The different treatment, which systematically varies with the experimental condition, may cause the participant to behave in a way that confirms the researcher's hypothesis or expectancy, making it impossible to determine whether it is the difference in the experimenter's behavior or in the experimental conditions that causes the change in the subject's behavior. Robert Rosenthal and his colleagues were among the first to establish experimenter expectancy effects when they told teachers that some of their students had been identified as "late bloomers" whose academic performance was expected to improve over the course of the school year. Although the students chosen to be designated as late bloomers had in fact been selected randomly, the teachers' expectations about their performance appeared to cause these students to improve.

### Response Bias

Another source of systematic error comes from participant response sets, such as the tendency for participants to answer questions in an agreeable manner (e.g., "yes" and "agree"), known as an *acquiescent response set*. If all the dependent measures are constructed such that agreement with an item means the same thing (e.g., agreement always indicates liberal political attitudes rather than sometimes indicating conservative attitudes and sometimes indicating liberal attitudes), it is possible for researchers to over- or underestimate the favorability of participants' attitudes, whether participants possess a particular trait, or the likelihood that they will engage in a particular behavior.

### Avoiding Bias

Careful research design can minimize systematic errors in collected data. Random sampling reduces sample bias. Random assignment to condition minimizes or eliminates selection bias. Ensuring that experimenters are blind to experimental conditions eliminates the possibility that experimenter expectancies will influence participant behavior or bias the data collected. Bias reduction improves researchers' ability to generalize findings and to draw causal conclusions from the data.

*Margaret Bull Kovera*

***See also*** Experimenter Expectancy Effect; Response Bias; Sampling; Selection; Systematic Error

#### Further Readings

Larzelere, R. E., Kuhn, B. R., & Johnson, B. (2004). The intervention selection bias: An unrecognized confound in intervention research. *Psychological Bulletin, 130,* 289–303.

Rosenthal, R. (2002). Covert communication in classrooms, clinics, courtrooms, and cubicles. *American Psychologist, 57,* 839–849.

Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research methods and data analysis* (Chapter 10, pp. 205–230). New York: McGraw-Hill.

Welkenhuysen-Gybels, J., Billiet, J., & Cambré, B. (2003). Adjustment for acquiescence in the assessment of the construct equivalence of Likert-type score items. *Journal of Cross-Cultural Psychology, 34,* 702–722.

## Biased Estimator

In many scientific research fields, statistical models are used to describe a system or a population, to interpret a phenomenon, or to investigate the

relationship among various measurements. These statistical models often contain one or multiple components, called *parameters,* that are unknown and thus need to be estimated from the data (sometimes also called the *sample*). An estimator, which is essentially a function of the observable data, is biased if its expectation does not equal the parameter to be estimated.

To formalize this concept, suppose $\theta$ is the parameter of interest in a statistical model. Let $\hat{\theta}$ be its estimator based on an observed sample. Then $\hat{\theta}$ is a biased estimator if $E(\hat{\theta}) \neq \theta$, where $E$ denotes the expectation operator. Similarly, one may say that $\hat{\theta}$ is an unbiased estimator if $E(\hat{\theta}) = \theta$. Some examples follow.

## Example 1

Suppose an investigator wants to know the average amount of credit card debt of undergraduate students from a certain university. Then the population would be all undergraduate students currently enrolled in this university, and the population mean of the amount of credit card debt of these undergraduate students, denoted by $\theta$, is the parameter of interest. To estimate $\theta$, a random sample is collected from the university, and the sample mean of the amount of credit card debt is calculated. Denote this sample mean by $\hat{\theta}_1$. Then $E(\hat{\theta}_1) = \theta$; that is, $\hat{\theta}_1$ is an unbiased estimator. If the largest amount of credit card debt from the sample, call it $\hat{\theta}_2$, is used to estimate $\theta$, then obviously $\hat{\theta}_2$ is biased. In other words, $E(\hat{\theta}_2) \neq \theta$.

## Example 2

In this example a more abstract scenario is examined. Consider a statistical model in which a random variable $X$ follows a normal distribution with mean $\mu$ and variance $\sigma^2$, and suppose a random sample $X_1, \ldots, X_n$ is observed. Let the parameter $\theta$ be $\mu$. It is seen in Example 1 that $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, the sample mean of $X_1, \ldots, X_n$, is an unbiased estimator for $\theta$. But $\overline{X}^2$ is a biased estimator for $\mu^2$ (or $\theta^2$). This is because $\overline{X}$ follows a normal distribution with mean $\mu$ and variance $\frac{\sigma^2}{n}$. Therefore, $E\left(\overline{X}^2\right) = \mu^2 + \frac{\sigma^2}{n} \neq \mu^2$.

Example 2 indicates that one should be careful about determining whether an estimator is biased. Specifically, although $\hat{\theta}$ is an unbiased estimator for $\theta$, $g(\hat{\theta})$ may be a biased estimator for $g(\theta)$ if $g$ is a nonlinear function. In Example 2, $g(\theta) = \theta^2$ is such a function. However, when $g$ is a linear function, that is, $g(\theta) = a\theta + b$ where $a$ and $b$ are two constants, then $g(\hat{\theta})$ is always an unbiased estimator for $g(\theta)$.

## Example 3

Let $X_1, \ldots, X_n$ be an observed sample from some distribution (not necessarily normal) with mean $\mu$ and variance $\sigma^2$. The sample variance $S^2$, which is defined as $\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$, is an unbiased estimator for $\sigma^2$, while the intuitive guess $\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ would yield a biased estimator. A heuristic argument is given here. If $\mu$ were known, $\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$ could be calculated, which would be an unbiased estimator for $\sigma^2$. But since $\mu$ is not known, it has to be replaced by $\overline{X}$. This replacement actually makes the numerator smaller. That is, $\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \leq \sum_{i=1}^{n}(X_i - \mu)^2$ regardless of the value of $\mu$. Therefore, the denominator has to be reduced a little bit (from $n$ to $n-1$) accordingly.

A closely related concept is the bias of an estimator, which is defined as $E(\hat{\theta}) - \theta$. Therefore, an unbiased estimator can also be defined as an estimator whose bias is zero, while a biased estimator is one whose bias is nonzero. A biased estimator is said to underestimate the parameter if the bias is negative or overestimate the parameter if the bias is positive.

Biased estimators are usually not preferred in estimation problems, because in the long run, they do not provide an accurate "guess" of the parameter. Sometimes, however, cleverly constructed biased estimators are useful because although their expectation does not equal the parameter under estimation, they may have a small

variance. To this end, a criterion that is quite commonly used in statistical science for judging the quality of an estimator needs to be introduced. The mean square error (MSE) of an estimator $\hat{\theta}$ for the parameter $\theta$ is defined as $E\left[\left(\hat{\theta}-\theta\right)^2\right]$. Apparently, one should seek estimators that make the MSE small, which means that $\hat{\theta}$ is "close" to $\theta$. Notice that

$$E\left[\left(\hat{\theta}-\theta\right)^2\right] = E\left[\left(\hat{\theta}-E\{\hat{\theta}\}\right)^2\right] + \left(E\{\hat{\theta}\}-\theta\right)^2 = Var\left(\hat{\theta}\right) + Bias^2,$$

meaning that the magnitude of the MSE, which is always nonnegative, is determined by two components: the variance and the bias of the estimator. Therefore, an unbiased estimator (for which the bias would be zero), if possessing a large variance, may be inferior to a biased estimator whose variance and bias are both small. One of the most prominent examples is the shrinkage estimator, in which a small amount of bias for the estimator gains a great reduction of variance. Example 4 is a more straightforward example of the usage of a biased estimator.

### Example 4

Let $X$ be a Poisson random variable, that is, $P(X=x)=\frac{e^{-\lambda}\lambda^x}{x!}$, for $x=0,1,2,\ldots$. Suppose the parameter $\theta=e^{-2\lambda}$, which is essentially $[P(X=0)]^2$, is of interest and needs to be estimated. If an unbiased estimator, say $\hat{\theta}_1(X)$, for $\theta$ is desired, then by the definition of unbiasedness, it must satisfy $\sum_{x=0}^{\infty}\hat{\theta}_1(x)\frac{e^{-\lambda}\lambda^x}{x!}=e^{-2\lambda}$ or, equivalently, $\sum_{x=0}^{\infty}\frac{\hat{\theta}_1(x)\lambda^x}{x!}=e^{-\lambda}$ for all positive values of $\lambda$. Clearly, the only solution is that $\hat{\theta}_1(x)=(-1)^x$. But this unbiased estimator is rather absurd. For example, if $X=10$, then the estimator $\hat{\theta}_1$ takes the value of 1, whereas if $X=11$, then $\hat{\theta}_1$ is $-1$. As a matter of fact, a much more reasonable estimator would be $\hat{\theta}_2(X)=e^{-2X}$, based on the maximum likelihood approach. This estimator is biased but always has a smaller MSE than $\hat{\theta}_1(X)$.

*Zhigang Zhang and Qianxing Mo*

***See also*** Distribution; Estimation; Expected Value

**Further Readings**

Rice, J. A. (1994). *Mathematical statistics and data analysis* (2nd ed.). Belmont, CA: Duxbury Press.

## Bivariate Regression

Regression is a statistical technique used to help investigate how variation in one or more variables predicts or explains variation in another variable. This popular statistical technique is flexible in that it can be used to analyze experimental or nonexperimental data with multiple categorical and continuous independent variables. If only one variable is used to predict or explain the variation in another variable, the technique is referred to as *bivariate regression*. When more than one variable is used to predict or explain variation in another variable, the technique is referred to as *multiple regression*. Bivariate regression is the focus of this entry.

Various terms are used to describe the independent variable in regression, namely, *predictor variable, explanatory variable,* or *presumed cause*. The dependent variable is often referred to as an *outcome variable, criterion variable,* or *presumed effect*. The choice of independent variable term will likely depend on the preference of the researcher or the purpose of the research. Bivariate regression may be used solely for predictive purposes. For example, do scores on a college entrance exam predict college grade point average? Or it may be used for explanation. Do differences in IQ scores explain differences in achievement scores? It is often the case that although the term *predictor* is used by researchers, the purpose of the research is, in fact, explanatory.

Suppose a researcher is interested in how well reading in first grade predicts or explains fifth-grade science achievement scores. The researcher hypothesizes that those who read well in first grade will also have high science achievement in fifth grade. An example bivariate regression will be performed to test this hypothesis. The data used in this example are a random sample of students (10%) with first-grade reading and fifth-grade

science scores and are taken from the Early Childhood Longitudinal Study public database. Variation in reading scores will be used to explain variation in science achievement scores, so first-grade reading achievement is the explanatory variable and fifth-grade science achievement is the outcome variable. Before the analysis is conducted, however, it should be noted that bivariate regression is rarely used in published research. For example, intelligence is likely an important common cause of both reading and science achievement. If a researcher was interested in explaining fifth-grade science achievement, then potential important common causes, such as intelligence, would need to be included in the research.

## Regression Equation

The simple equation for bivariate linear regression is $Y = a + bX + e$. The science achievement score, $Y$, for a student equals the intercept or constant ($a$), plus the slope ($b$) times the reading score ($X$) for that student, plus error ($e$). Error, or the residual component ($e$), represents the error in prediction, or what is not explained in the outcome variable. The error term is not necessary and may be dropped so that the following equation is used: $Y' = a + bX$. $Y'$ is the expected (or predicted) score. The intercept is the predicted fifth-grade science score for someone whose first-grade reading score is zero. The slope ($b$, also referred to as the *unstandardized regression coefficient*) represents the predicted unit increase in science scores associated with a one-unit increase in reading scores. $X$ is the observed score for that person. The two parameters ($a$ and $b$) that describe the linear relation between the predictor and outcome are thus the intercept and the regression coefficient. These parameters are often referred to as least squares estimators and will be estimated using the two sets of scores. That is, they represent the optimal estimates that will provide the least error in prediction.

Returning to the example, the data used in the analysis were first-grade reading scores and fifth-grade science scores obtained from a sample of 1,027 school-age children. $T$-scores, which have a mean of 50 and standard deviation of 10, were used. The means for the scores in the sample were 51.31 for reading and 51.83 for science.

Because science scores are the outcome, the science scores are regressed on first-grade reading scores. The easiest way to conduct such analysis is to use a statistical program. The estimates from the output may then be plugged into the equation. For these data, the prediction equation is $Y' = 21.99 + (.58)X$. Therefore, if a student's first-grade reading score was 60, the predicted fifth-grade science achievement score for that student would be $21.99 + (.58)60$, which equals 56.79. One might ask, why even conduct a regression analysis to obtain a predicted science score when Johnny's science score was already available? There are a few possible reasons. First, perhaps a researcher wants to use the information to predict later science performance, either for a new group of students or for an individual student, based on current first-grade reading scores. Second, a researcher may want to know the relation between the two variables, and a regression provides a nice summary of the relation between the scores for all the students. For example, do those students who tend to do well in reading in first grade also do well in science in fifth grade? Last, a researcher might be interested in different outcomes related to early reading ability when considering the possibility of implementing an early reading intervention program. Of course a bivariate relation is not very informative. A much more thoughtfully developed causal model would need to be developed if a researcher was serious about this type of research.

## Scatterplot and Regression Line

The regression equation describes the linear relation between variables; more specifically, it describes science scores as a function of reading scores. A scatterplot could be used to represent the relation between these two variables, and the use of a scatterplot may assist one in understanding regression. In a scatterplot, the science scores (outcome variable) are on the $y$-axis, and the reading scores (explanatory variable) are on the $x$-axis.

A scatterplot is shown in Figure 1. Each person's reading and science scores in the sample are plotted. The scores are clustered fairly closely together, and the general direction looks to be positive. Higher scores in reading are generally associated with higher scores in science. The next step is

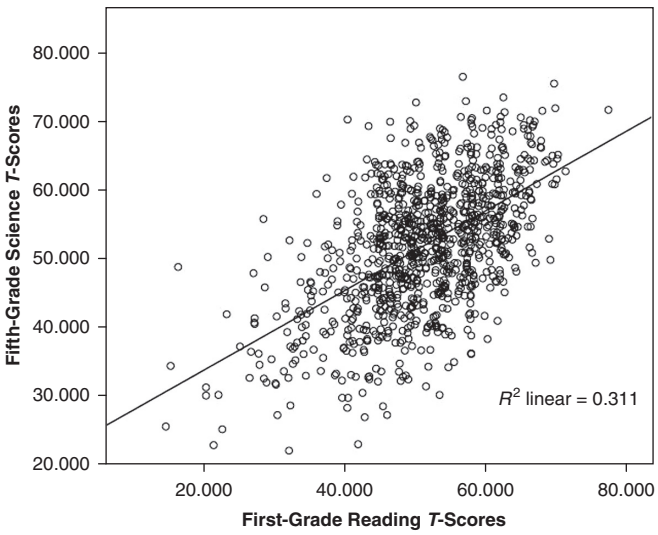



**Figure 1** Scatterplot and Regression Line

to fit a regression line. The regression line is plotted so that it minimizes errors in prediction, or simply, the regression line is the line that is closest to all the data points. The line is fitted automatically in many computer programs, but information obtained in the regression analysis output can also be used to plot two data points that the line should be drawn through. For example, the intercept (where the line crosses the *y*-axis) represents the predicted science score when reading equals zero. Because the value of the intercept was 21.99, the first data point would be found at 0 on the *x*-axis and at 21.99 on the *y*-axis. The second point on the line may be located at the mean reading score (51.31) and mean science score (51.83). A line can then be drawn through those two points. The line is shown in Figure 1. Points that are found along this regression line represent the predicted science achievement score for Person A with a reading score of X.

## Unstandardized and Standardized Coefficients

For a more thorough understanding of bivariate regression, it is useful to examine in more detail the output obtained after running the regression. First, the intercept has no important substantive meaning. It is unlikely that anyone would score a zero on the reading test, so it does not make much sense. It is useful in the unstandardized solution in that it is used to obtain predicted scores (it is a constant value added to everyone's score), and as demonstrated above, it is useful in plotting a regression line. The slope ($b = .58$) is the unstandardized coefficient. It was statistically significant, indicating that reading has a statistically significant influence on fifth-grade science. A 1-point *T*-score increase in reading is associated with a .58 *T*-score point increase in science scores. The *b*s are interpreted in the metric of the original variable. In the example, all the scores were *T*-scores. Unstandardized coefficients are especially useful for interpretation when the metric of the variables is meaningful. Sometimes, however, the metric of the variables is not meaningful.

Two equations were generated in the regression analysis. The first, as discussed in the example above, is referred to as the *unstandardized solution*. In addition to the unstandardized solution, there is a *standardized solution*. In this equation, the constant is dropped, and *z* scores (mean = 0, standard deviation = 1), rather than the *T*-scores (or raw scores), are used. The standardized regression coefficient is referred to as a *beta weight* ($\beta$). In the example, the beta weight was .56. Therefore, a one-standard-deviation increase in reading was associated with a .56-standard-deviation increase in science achievement. The unstandardized and standardized coefficients were similar in this example because *T*-scores are standardized scores, and the sample statistics for the *T*-scores were fairly close to the population mean of 50 and standard deviation of 10.

It is easy to convert back and forth from standardized to unstandardized regression coefficients:

$$\beta = b\left(\frac{\text{standard deviation of reading scores}}{\text{standard deviation of science scores}}\right)$$

or

$$b = \beta\left(\frac{\text{standard deviation of science scores}}{\text{standard deviation of reading scores}}\right).$$

From an interpretative standpoint, should someone interpret the unstandardized or the standardized coefficient? There is some debate over which

one to use for interpretative statements, but in a bivariate regression, the easiest answer is that if both variables are in metrics that are easily interpretable, then it would make sense to use the unstandardized coefficients. If the metrics are not meaningful, then it may make more sense to use the standardized coefficient. Take, for example, number of books read per week. If number of books read per week was represented by the actual number of books read per week, the variable is in a meaningful metric. If the number of books read per week variable were coded so that 0 = no books read per week, 1 = one to three books read per week, and 2 = four or more books read per week, then the variable is not coded in a meaningful metric, and the standardized coefficient would be the better one to use for interpretation.

## *R* and $R^2$

In bivariate regression, typically the regression coefficient is of greatest interest. Additional information is provided in the output, however. *R* is used in multiple regression output and represents a multiple correlation. Because there is only one explanatory variable, *R* (.56) is equal to the correlation coefficient ($r$ = .56) between reading and science scores. Note that this value is also identical to the $\beta$. Although the values of $\beta$ and $r$ are the same, the interpretation differs. The researcher is not proposing an agnostic relation between reading scores and science scores. Rather the researcher is positing that early reading explains later science achievement. Hence, there is a clear direction in the relation, and this direction is not specified in a correlation.

$R^2$ is the variance in science scores explained by reading scores. In the current example, $R^2 = 31$. First-grade reading scores explained 31% of the variance in fifth-grade science achievement scores.

## Statistical Significance

*R* and $R^2$ are typically used to evaluate the statistical significance of the overall regression equation (the tests for the two will result in the same answer). The null hypothesis is that $R^2$ equals zero in the population. One way of calculating the statistical significance of the overall regression is to use an *F* test associated with the value obtained with the formula

$$\frac{R^2/k}{(1-R^2)/(N-k-1)}.$$

In this formula, $R^2$ equals the variance explained, $1 - R^2$ is the variance unexplained, and $k$ equals the degrees of freedom (*df*) for the regression (which is 1 because one explanatory variable was used). With the numbers plugged in, the formula would look like

$$\frac{.31/1}{.69/(1027-1-1)}$$

and results in $F = 462.17$. An *F* table indicates that reading did have a statistically significant effect on science achievement, $R^2 = .31$, $F(1,1025) = 462.17, p < .01$.

In standard multiple regression, a researcher typically interprets the statistical significance of $R^2$ (the statistical significance of the overall equation) and the statistical significance of the unique effects of each individual explanatory variable. Because this is bivariate regression, however, the statistical significance test of the overall regression and the regression coefficient (*b*) will yield the same results, and typically the statistical significance tests for each are not reported.

The statistical significance of the regression coefficient (*b*) is evaluated with a *t* test. The null hypothesis is that the slope equals zero, that is, the regression line is parallel with the *x*-axis. The *t*-value is obtained by

$$\frac{b}{\text{standard error of } b}.$$

In this example, $b = .58$, and its associated standard error was .027. The *t*-value was 21.50. A *t*-table could be consulted to determine whether 21.50 is statistically significant. Or a rule of thumb may be used that given the large sample size and with a two-tailed significance test, a *t*-value greater than 2 will be statistically significant at the $p < .05$ level. Clearly, the regression coefficient was statistically significant. Earlier it was mentioned that because this is a bivariate regression, the significance of the overall regression and *b* provide redundant information. The use of *F* and *t* tests may thus

be confusing, but note that $F$ (462.17) equals $t^2$ ($21.50^2$) in this bivariate case. A word of caution: This finding does not generalize to multiple regression. In fact, in a multiple regression, the overall regression might be significant, and some of the $b$s may or may not be significant. In a multiple regression, both the overall regression equation and the individual coefficients are examined for statistical significance.

## Residuals

Before completing this explanation of bivariate regression, it will be instructive to discuss a topic that has been for the most part avoided until now: the residuals. Earlier it was mentioned that $e$ (residual) was also included in the regression equation. Remember that regression parameter estimates minimize the prediction errors, but the prediction is unlikely to be perfect. The residuals represent the error in prediction. Or the residual variance represents the variance that is left unexplained by the explanatory variable. Returning to the example, if reading scores were used to predict science scores for those 1,026 students, each student would have a prediction equation in which his or her reading score would be used to calculate a predicted science score. Because the actual score for each person is also known, the residual for each person would represent the observed fifth-grade science score minus the predicted score obtained from the regression equation. Residuals are thus observed scores minus predicted scores, or conceptually they may be thought of as the fifth-grade science scores with effects of first-grade reading removed.

Another way to think of the residuals is to revert back to the scatterplot in Figure 1. The $x$-axis represents the observed scores, and the $y$-axis represents the science scores. Both predicted and actual scores are already plotted on this scatterplot. That is, the predicted scores are found on the regression line. If a person's reading score was 40, the predicted science score may be obtained by first finding 40 on the $x$-axis, and then moving up in a straight line until reaching the regression line. The observed science scores for this sample are also shown on the plot, represented by the dots scattered about the regression line. Some are very close to the line whereas others are farther away. Each person's residual is thus represented by the distance between the observed score and the regression line. Because the regression line represents the predicted scores, the residuals are the difference between predicted and observed scores. Again, the regression line minimizes the distance of these residuals from the regression line. Much as residuals are thought of as science scores with the effects of reading scores removed, the residual variance is the proportion of variance in science scores left unexplained by reading scores. In the example, the residual variance was .69, or $1 - R^2$.

## Regression Interpretation

An example interpretation for the reading and science example concludes this entry on bivariate regression. The purpose of this study was to determine how well first-grade science scores explained fifth-grade science achievement scores. The regression of fifth-grade science scores on first-grade reading scores was statistically significant, $R^2 = .31,\ F(1, 1025) = 462.17, p < .01$. Reading accounted for 31% of the variance in science achievement. The unstandardized regression coefficient was .58, meaning that for each $T$-score point increase in reading, there was a .58 $T$-score increase in science achievement. Children who are better readers in first grade also tend to be higher achievers in fifth-grade science.

*Matthew R. Reynolds*

***See also*** Correlation; Multiple Regression; Path Analysis; Scatterplot; Variance

### Further Readings

Bobko, P. (2001). *Correlation and regression: Principles and applications for industrial/organizational psychology and management* (2nd ed.). Thousand Oaks, CA: Sage.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Keith, T. Z. (2006). *Multiple regression and beyond.* Boston: Pearson.

McDonald, R. P. (1999). *Test theory: A unified treatment.* Mahwah, NJ: Lawrence Erlbaum.

Miles, J., & Shevlin, M. (2001). *Applying regression and correlation: A guide for students and researchers*. Thousand Oaks, CA: Sage.
Schroeder, L. D., Sjoquist, D. L., & Stephan, P. E. (1986). *Understanding regression analysis: An introductory guide*. Beverly Hills, CA: Sage.
Weisburg, S. (2005). *Applied linear regression* (3rd ed.). Hoboken, NJ: Wiley.

# Block Design

Sir Ronald Fisher, the father of modern experimental design, extolled the advantages of block designs in his classic book, *The Design of Experiments*. He observed that block designs enable researchers to reduce error variation and thereby obtain more powerful tests of false null hypotheses. In the behavioral sciences, a significant source of error variation is the nuisance variable of individual differences. This nuisance variable can be isolated by assigning participants or experimental units to blocks so that at the beginning of an experiment, the participants within a block are more homogeneous with respect to the dependent variable than are participants in different blocks. Three procedures are used to form homogeneous blocks.

1. Match participants on a variable that is correlated with the dependent variable. Each block consists of a set of matched participants.
2. Observe each participant under all or a portion of the treatment levels or treatment combinations. Each block consists of a single participant who is observed two or more times. Depending on the nature of the treatment, a period of time between treatment level administrations may be necessary in order for the effects of one treatment level to dissipate before the participant is observed under other levels.
3. Use identical twins or litter mates. Each block consists of participants who have identical or similar genetic characteristics.

Block designs also can be used to isolate other nuisance variables, such as the effects of administering treatments at different times of day, on different days of the week, or in different testing facilities. The salient features of the five most often used block designs are described next.

## Block Designs With One Treatment

### *Dependent Samples* t-*Statistic Design*

The simplest block design is the randomization and analysis plan that is used with a $t$ statistic for dependent samples. Consider an experiment to compare two ways of memorizing Spanish vocabulary. The dependent variable is the number of trials required to learn the vocabulary list to the criterion of three correct recitations. The null and alternative hypotheses for the experiment are, respectively,

$$H_0: \mu_1 - \mu_2 = 0$$

and

$$H_1: \mu_1 - \mu_2 \neq 0,$$

where $\mu_1$ and $\mu_2$ denote the population means for the two memorization approaches. It is reasonable to believe that IQ is negatively correlated with the number of trials required to memorize Spanish vocabulary. To isolate this nuisance variable, $n$ blocks of participants can be formed so that the two participants in each block have similar IQs. A simple way to form blocks of matched participants is to rank the participants in terms of IQ. The participants ranked 1 and 2 are assigned to Block 1, those ranked 3 and 4 are assigned to Block 2, and so on. Suppose that 20 participants have volunteered for the memorization experiment. In this case, $n = 10$ blocks of dependent samples can be formed. The two participants in each block are randomly assigned to the memorization approaches. The layout for the experiment is shown in Figure 1.

The null hypothesis is tested using a $t$ statistic for dependent samples. If the researcher's hunch is correct—that IQ is correlated with the number of trials to learn—the design should result in a more powerful test of a false null hypothesis than would a $t$-statistic design for independent samples. The increased power results from isolating the nuisance variable of IQ so that it does not appear in the estimates of the error effects.

| | Treat. Level | Dep. Var. | Treat. Level | Dep. Var. |
|---|---|---|---|---|
| $Block_1$ | $a_1$ | $Y_{11}$ | $a_2$ | $Y_{12}$ |
| $Block_2$ | $a_1$ | $Y_{21}$ | $a_2$ | $Y_{22}$ |
| $Block_3$ | $a_1$ | $Y_{31}$ | $a_2$ | $Y_{32}$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| $Block_{10}$ | $a_1$ | $Y_{10,1}$ | $a_2$ | $Y_{10,2}$ |
| | | $\overline{Y}_{\cdot 1}$ | | $\overline{Y}_{\cdot 2}$ |

**Figure 1** Layout for a Dependent Samples *t*-Statistic Design

*Notes:* $a_j$ denotes a treatment level (*Treat. Level*); $Y_{ij}$ denotes a measure of the dependent variable (Dep. Var.). Each block in the memorization experiment contains two matched participants. The participants in each block are randomly assigned to the treatment levels. The means of the treatments levels are denoted by $\overline{Y}_{\cdot 1}$ and $\overline{Y}_{\cdot 2}$.

### *Randomized Block Design*

The randomized block analysis of variance design can be thought of as an extension of a dependent samples *t*-statistic design for the case in which the treatment has two or more levels. The layout for a randomized block design with $p = 3$ levels of Treatment $A$ and $n = 10$ blocks is shown in Figure 2. A comparison of this layout with that in Figure 1 for the dependent samples *t*-statistic design reveals that the layouts are the same except that the randomized block design has three treatment levels.

In a randomized block design, a block might contain a single participant who is observed under all $p$ treatment levels or $p$ participants who are similar with respect to a variable that is correlated with the dependent variable. If each block contains one participant, the order in which the treatment levels are administered is randomized independently for each block, assuming that the nature of the research hypothesis permits this. If a block contains $p$ matched participants, the participants in each block are randomly assigned to the treatment levels.

The statistical analysis of the data is the same whether repeated measures or matched participants are used. However, the procedure used to form homogeneous blocks does affect the interpretation of the results. The results of an experiment with repeated measures generalize to a population of participants who have been exposed to all the treatment levels. The results of an experiment with matched participants generalize to a population of participants who have been exposed to only one treatment level.

The total sum of squares (*SS*) and total degrees of freedom for a randomized block design are partitioned as follows:

$$\begin{aligned} SSTOTAL &= SSA + SSBLOCKS + SSRESIDUAL \\ np - 1 &= (p-1) + \quad (n-1) \quad + (n-1)(p-1), \end{aligned}$$

where *SSA* denotes the Treatment *A SS* and *SSBLOCKS* denotes the blocks *SS*. The *SSRESIDUAL* is the interaction between Treatment *A* and blocks; it is used to estimate error effects. Many test statistics can be thought of as a ratio of error effects and treatment effects as follows:

$$\text{Test statistic} = \frac{f(\text{error effects}) + f(\text{treatment effects})}{f(\text{error effects})},$$

where $f()$ denotes a function of the effects in parentheses. The use of a block design enables a researcher to isolate variation attributable to the blocks variable so that it does not appear in estimates of error effects. By removing this nuisance variable from the numerator and denominator of the test statistic, a researcher is rewarded with a more powerful test of a false null hypothesis.

Two null hypotheses can be tested in a randomized block design. One hypothesis concerns the equality of the Treatment *A* population means; the other hypothesis concerns the equality of the blocks population means. For this design and those described later, assume that the treatment represents a fixed effect and the nuisance variable, blocks, represents a random effect. For this mixed model, the null hypotheses are

$H_0$: $\mu_{\cdot 1} = \mu_{\cdot 2} = \cdots = \mu_{\cdot p}$

(treatment *A* population means are equal)

| | Treat. Level | Dep. Var. | Treat. Level | Dep. Var. | Treat. Level | Dep. Var. | |
|---|---|---|---|---|---|---|---|
| $\text{Block}_1$ | $a_1$ | $Y_{11}$ | $a_2$ | $Y_{12}$ | $a_3$ | $Y_{13}$ | $\overline{Y}_{1\cdot}$ |
| $\text{Block}_2$ | $a_1$ | $Y_{21}$ | $a_2$ | $Y_{22}$ | $a_3$ | $Y_{23}$ | $\overline{Y}_{2\cdot}$ |
| $\text{Block}_3$ | $a_1$ | $Y_{31}$ | $a_2$ | $Y_{32}$ | $a_3$ | $Y_{33}$ | $\overline{Y}_{3\cdot}$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| $\text{Block}_{10}$ | $a_1$ | $Y_{10,1}$ | $a_2$ | $Y_{10,2}$ | $a_3$ | $Y_{10,3}$ | $\overline{Y}_{10\cdot}$ |
| | | $\overline{Y}_{\cdot 1}$ | | $\overline{Y}_{\cdot 2}$ | | $\overline{Y}_{\cdot 3}$ | |

**Figure 2** Layout for a Randomized Block Design With $p = 3$ Treatment Levels and $n = 10$ Blocks

*Notes:* $a_j$ denotes a treatment level (Treat. Level); $Y_{ij}$ denotes a measure of the dependent variable (Dep. Var.). Each block contains three matched participants. The participants in each block are randomly assigned to the treatment levels. The means of treatment $A$ are denoted by and $\overline{Y}_{\cdot 1}$ and $\overline{Y}_{\cdot 2}$ and $\overline{Y}_{\cdot 3}$ and the means of the blocks are denoted by $\overline{Y}_{\cdot 1}, \ldots, \overline{Y}_{10\cdot}$.

$$H_0: \sigma^2_{BL} = 0$$

(variance of the blocks, BL, population means is equalto zero)

where $\mu_{ij}$ denotes the population mean for the $i$th block and the $j$th level of treatment $A$.

The $F$ statistics for testing these hypotheses are

$$F = \frac{SSA/(p-1)}{SSRESIDUAL/[(n-1)(p-1)]} = \frac{MSA}{MSRESIDUAL}$$

and

$$F = \frac{SSBL/(n-1)}{SSRESIDUAL/[(n-1)(p-1)]} = \frac{MSBL}{MSRESIDUAL}.$$

The test of the blocks null hypothesis is generally of little interest because the population means of the nuisance variable are expected to differ.

The advantages of the design are simplicity in the statistical analysis and the ability to isolate a nuisance variable so as to obtain greater power to reject a false null hypothesis. The disadvantages of the design include the difficulty of forming homogeneous blocks and of observing participants $p$ times when $p$ is large and the restrictive sphericity assumption of the design. This assumption states that in order for $F$ statistics to be distributed as central $F$ when the null hypothesis is true, the variances of the differences for all pairs of treatment levels must be homogeneous; that is,

$$\sigma^2_{Y_j - Y_{j'}} = \sigma^2_j + \sigma^2_{j'} - 2\sigma_{jj'} \text{ for all } j \text{ and } j'.$$

### *Generalized Randomized Block Design*

A generalized randomized block design is a variation of a randomized block design. Instead of having $n$ blocks of $p$ homogeneous participants, the generalized randomized block design has $w$ groups of $np$ homogeneous participants. The $w$ groups, like the $n$ blocks in a randomized block design, represent a nuisance variable that a researcher wants to remove from the error effects. The generalized randomized block design can be used when a researcher is interested in one treatment with $p \geq 2$ treatment levels and the researcher has sufficient participants to form $w$ groups, each containing $np$ homogeneous participants. The total number of participants in the design is $N = npw$. The layout for the design is shown in Figure 3.

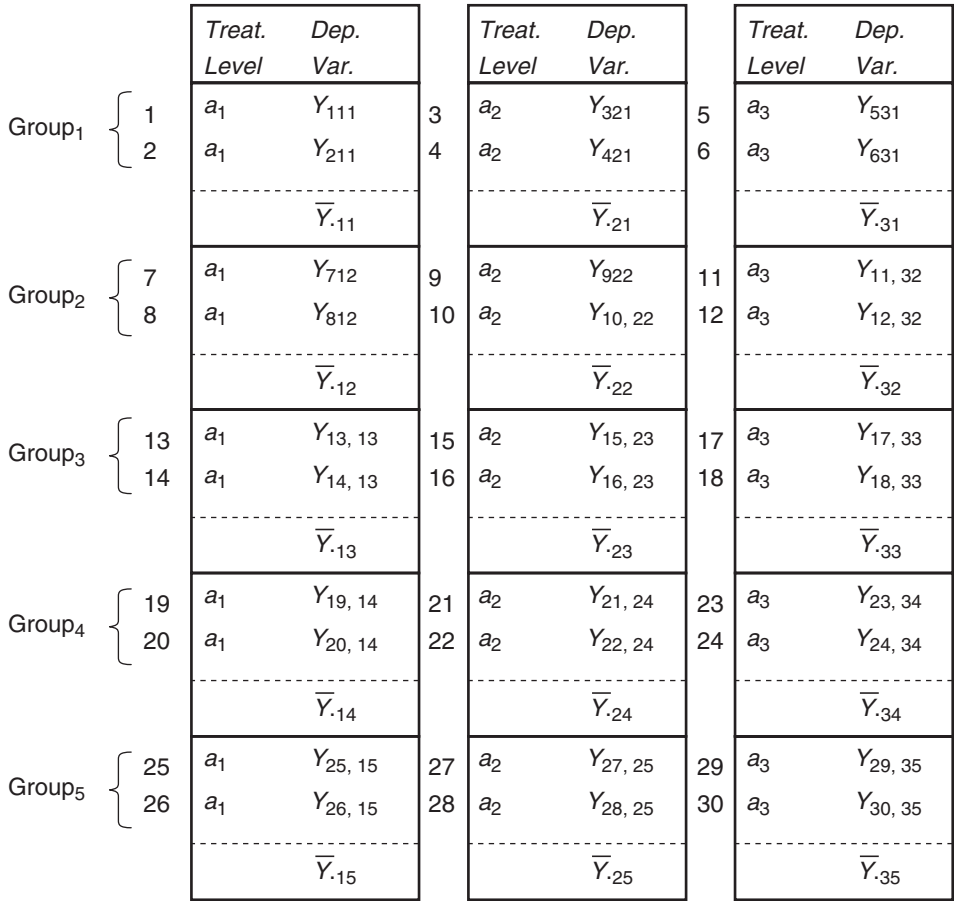



**Figure 3** Generalized Randomized Block Design With $N = 30$ Participants, $p = 3$ Treatment Levels, and $w = 5$ Groups of $np = (2)(3) = 6$ Homogeneous Participants

In the memorization experiment described earlier, suppose that 30 volunteers are available. The 30 participants are ranked with respect to IQ. The $np = (2)(3) = 6$ participants with the highest IQs are assigned to Group 1, the next 6 participants are assigned to Group 2, and so on. The $np = 6$ participants in each group are then randomly assigned to the $p = 3$ treatment levels with the restriction that $n = 2$ participants are assigned to each level.

The total $SS$ and total degrees of freedom are partitioned as follows:

$$SSTOTAL = SSA + SSG + SSA \times G + SSWCELL$$
$$npw - 1 = (p-1) + (w-1) + (p-1)(w-1) + pw(n-1),$$

where $SSG$ denotes the groups $SS$ and $SSA \times G$ denotes the interaction of Treatment $A$ and groups. The within-cells $SS, SSWCELL$, is used to estimate error effects. Three null hypotheses can be tested:

1. $H_0: \mu_{1\cdot} = \mu_{2\cdot} = \cdots = \mu_{p\cdot}$
(Treatment $A$ population means are equal),

2. $H_0: \sigma^2_G = 0$
(Variance of the groups, $G$, population means is equal to zero),

3. $H_0: \sigma^2_{A \times G} = 0$
(Variance of the $A \times G$ interaction is equal to zero),

where $\mu_{ijz}$ denotes a population mean for the $i$th participant in the $j$th treatment level and $z$th group.

| | Treat. Comb. | Dep. Var. | Treat. Comb. | Dep. Var. | Treat. Comb. | Dep. Var. | Treat. Comb. | Dep. Var. | |
|---|---|---|---|---|---|---|---|---|---|
| $\text{Block}_1$ | $a_1b_1$ | $Y_{111}$ | $a_1b_2$ | $Y_{112}$ | $a_2b_1$ | $Y_{121}$ | $a_2b_2$ | $Y_{122}$ | $\overline{Y}_{1..}$ |
| $\text{Block}_2$ | $a_1b_1$ | $Y_{211}$ | $a_1b_2$ | $Y_{212}$ | $a_2b_1$ | $Y_{221}$ | $a_2b_2$ | $Y_{222}$ | $\overline{Y}_{2..}$ |
| $\text{Block}_3$ | $a_1b_1$ | $Y_{311}$ | $a_1b_2$ | $Y_{312}$ | $a_2b_1$ | $Y_{321}$ | $a_2b_2$ | $Y_{322}$ | $\overline{Y}_{3..}$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| $\text{Block}_{10}$ | $a_1b_1$ | $Y_{10,\,11}$ | $a_1b_2$ | $Y_{10,\,12}$ | $a_2b_1$ | $Y_{10,\,21}$ | $a_2b_2$ | $Y_{10,\,22}$ | $\overline{Y}_{10..}$ |
| | | $\overline{Y}_{\cdot 11}$ | | $\overline{Y}_{\cdot 12}$ | | $\overline{Y}_{\cdot 21}$ | | $\overline{Y}_{\cdot 22}$ | |

**Figure 4** Layout for a Two-Treatment, Randomized Block Factorial Design in Which Four Homogeneous Participants Are Randomly Assigned to the $pq = 2 \times 2 = 4$ Treatment Combinations in Each Block

*Note:* $a_jb_k$ denotes a treatment combination (Treat. Comb.); $Y_{ijk}$ denotes a measure of the dependent variable (Dep. Var.).

The three null hypotheses are tested using the following $F$ statistics:

1. $F = \dfrac{SSA/(p-1)}{SSWCELL/[pw(n-1)]} = \dfrac{MSA}{MSWCELL}$,
2. $F = \dfrac{SSG/(w-1)}{SSWCELL/[pw(n-1)]} = \dfrac{MSG}{MSWCELL}$,
3. $F = \dfrac{SSA \times G/(p-1)(w-1)}{SSWCELL/[pw(n-1)]} = \dfrac{MSA \times G}{MSWCELL}$.

The generalized randomized block design enables a researcher to isolate one nuisance variable—an advantage that it shares with the randomized block design. Furthermore, the design uses the within-cell variation in the $pw = (3)(5) = 15$ cells to estimate error effects rather than an interaction, as in the randomized block design. Hence, the restrictive sphericity assumption of the randomized block design is replaced with the assumption of homogeneity of within-cell population variances.

## Block Designs With Two or More Treatments

The blocking procedure that is used with a randomized block design can be extended to experiments that have two or more treatments, denoted by the letters $A$, $B$, $C$, and so on.

### *Randomized Block Factorial Design*

A randomized block factorial design with two treatments, denoted by $A$ and $B$, is constructed by crossing the $p$ levels of Treatment $A$ with the $q$ levels of Treatment $B$. The design's $n$ blocks each contain $p \times q$ treatment combinations: $a_1b_1, a_1b_2 \ldots a_pb_q$. The design enables a researcher to isolate variation attributable to one nuisance variable while simultaneously evaluating two treatments and associated interaction.

The layout for the design with $p = 2$ levels of Treatment $A$ and $q = 2$ levels of Treatment $B$ is shown in Figure 4. It is apparent from Figure 4 that all the participants are used in simultaneously evaluating the effects of each treatment. Hence, the design permits efficient use of resources because each treatment is evaluated with the same precision as if the entire experiment had been devoted to that treatment alone.

The total $SS$ and total degrees of freedom for a two-treatment randomized block factorial design are partitioned as follows:

$$SSTOTAL = SSBL + SSA + SSB + SSA \times B + SSRESIDUAL$$
$$npq - 1 = (n-1) + (p-1) + (q-1) + (p-1)(q-1) + (n-1)(pq-1).$$

Four null hypotheses can be tested:

1. $H_0: \sigma^2_{BL} = 0$ (Variance of the blocks, $BL$, population means is equal to zero),
2. $H_0: \mu_{.1.} = \mu_{.2.} = \cdots = \mu_{.p.}$ (Treatment $A$ population means are equal),
3. $H_0: \mu_{..1} = \mu_{..2} = \cdots = \mu_{..q}$ (Treatment $B$ population means are equal),
4. $H_0: A \times B$ interaction = 0 (Treatments $A$ and $B$ do not interact),

where $\mu_{ijk}$ denotes a population mean for the $i$th block, $j$th level of treatment $A$, and $k$th level of treatment $B$. The $F$ statistics for testing the null hypotheses are as follows:

$$F = \frac{SSBL/(n-1)}{SSRESIDUAL/[(n-1)(pq-1)]} = \frac{MSBL}{MSRESIDUAL},$$

$$F = \frac{SSA/(p-1)}{SSRESIDUAL/[(n-1)(pq-1)]} = \frac{MSA}{MSRESIDUAL},$$

$$F = \frac{SSB/(q-1)}{SSRESIDUAL/[(n-1)(pq-1)]} = \frac{MSB}{MSRESIDUAL},$$

$$F = \frac{SSA \times B/(p-1)(q-1)}{SSRESIDUAL/[(n-1)(pq-1)]} = \frac{MSA \times B}{MSRESIDUAL}.$$

The design shares the advantages and disadvantages of the randomized block design. Furthermore, the design enables a researcher to efficiently evaluate two or more treatments and associated interactions in the same experiment. Unfortunately, the design lacks simplicity in the interpretation of the results if interaction effects are present. The design has another disadvantage: If Treatment $A$ or $B$ has numerous levels, say four or five, the block size becomes prohibitively large. For example, if $p = 4$ and $q = 3$, the design has blocks of size $4 \times 3 = 12$. Obtaining $n$ blocks with 12 matched participants or observing $n$ participants on 12 occasions is often not feasible. A design that reduces the size of the blocks is described next.

### *Split-Plot Factorial Design*

In the late 1920s, Fisher and Frank Yates addressed the problem of prohibitively large block sizes by developing confounding schemes in which only a portion of the treatment combinations in an experiment are assigned to each block. The split-plot factorial design achieves a reduction in the block size by confounding one or more treatments with groups of blocks. *Group–treatment confounding* occurs when the effects of, say, Treatment $A$ with $p$ levels are indistinguishable from the effects of $p$ groups of blocks.

The layout for a two-treatment split-plot factorial design is shown in Figure 5. The block size in the split-plot factorial design is half as large as the block size of the randomized block factorial design in Figure 4 although the designs contain the same treatment combinations. Consider the sample means $\overline{Y}_{.1.}$ and $\overline{Y}_{.2.}$ in Figure 5. Because of confounding, the difference between $\overline{Y}_{.1.}$ and $\overline{Y}_{.2.}$ reflects both group effects and Treatment $A$ effects.

The total $SS$ and total degrees of freedom for a split-plot factorial design are partitioned as follows:

$$\begin{aligned} SSTOTAL &= SSA + SSBL(A) + SSB + SSA \times B + SSRESIDUAL \\ npq - 1 &= (p-1) + p(n-1) + (q-1) + (p-1)(q-1) + p(n-1)(q-1), \end{aligned}$$

where $SSBL(A)$ denotes the $SS$ for blocks within Treatment $A$. Three null hypotheses can be tested:

1. $H_0: \mu_{.1.} = \mu_{.2.} = \cdots = \mu_{.p}$ (Treatment $A$ population means are equal),
2. $H_0: \mu_{..1} = \mu_{..2} = \cdots = \mu_{..q}$ (Treatment $B$ population means are equal),
3. $H_0: A \times B$ interaction = 0 (Treatments $A$ and $B$ do not interact),

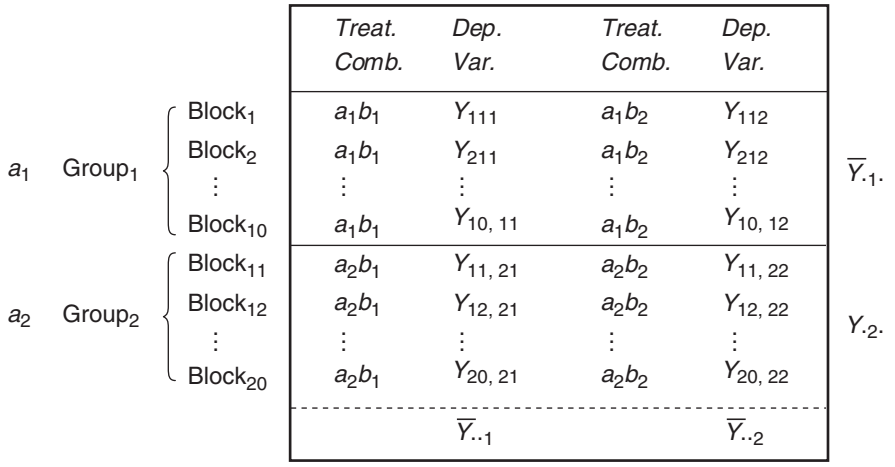



**Figure 5** Layout for a Two-Treatment, Split-Plot Factorial Design in Which 10 + 10 = 20 Homogeneous Blocks Are Randomly Assigned to the Two Groups

*Notes:* $a_jb_k$ denotes a treatment combination (Treat. Comb.); $Y_{ijk}$ denotes a measure of the dependent variable (Dep. Var.). Treatment $A$ is confounded with groups. Treatment $B$ and the $A \times B$ are not confounded.

where $\mu_{ijk}$ denotes the $i$th block, $j$th level of treatment $A$, and $k$th level of treatment $B$. The $F$ statistics are

$$F = \frac{SSA/(p-1)}{SSBL(A)/[p(n-1)]} = \frac{MSA}{MSBL(A)},$$

$$F = \frac{SSB/(q-1)}{SSRESIDUAL/[p(n-1)(q-1)]}$$

$$= \frac{MSB}{MSRESIDUAL},$$

$$F = \frac{SSA \times B/(p-1)(q-1)}{SSRESIDUAL/[p(n-1)(q-1)]}$$

$$= \frac{MSA \times B}{MSRESIDUAL}.$$

The split-plot factorial design uses two error terms: $MSBL(A)$ is used to test Treatment $A$; a different and usually much smaller error term, $MSRESIDUAL$, is used to test Treatment $B$ and the $A \times B$ interaction. Because $MSRESIDUAL$ is generally smaller than $MSBL(A)$, the power of the tests of Treatment $B$ and the $A \times B$ interaction is greater than that for Treatment $A$.

*Roger E. Kirk*

***See also*** Analysis of Variance (ANOVA); Confounding; *F* Test; Nuisance Variable; Null Hypothesis; Sphericity; Sums of Squares

**Further Readings**

Dean, A., & Voss, D. (1999). *Design and analysis of experiments*. New York: Springer-Verlag.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole.

Kirk, R. E. (2002). Experimental design. In I. B. Weiner (Series Ed.) & J. Schinka & W. F. Velicer (Vol. Eds.), *Handbook of psychology: Vol. 2. Research methods in psychology* (pp. 3–32). New York: Wiley.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Myers, J. L., & Well, A. D. (2003). *Research design and statistical analysis* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

# Bonferroni Procedure

The Bonferroni procedure is a statistical adjustment to the significance level of hypothesis tests when multiple tests are being performed. The purpose of an adjustment such as the Bonferroni procedure is to reduce the probability of identifying significant results that do not exist, that is, to guard against making Type I errors (rejecting null hypotheses when they are true) in the testing process. This potential for error increases with an increase in the number of tests being performed in a given study and is due to the multiplication of probabilities across the multiple tests. The Bonferroni procedure is often used as an adjustment in multiple comparisons after a significant finding in an analysis of variance (ANOVA) or when constructing simultaneous confidence intervals for several population parameters, but more broadly, it can be used in any situation that involves multiple tests. The Bonferroni procedure is one of the more commonly used procedures in multiple testing situations, primarily because it is an easy adjustment to make. A strength of the Bonferroni procedure is its ability to maintain Type I error rates at or below a nominal value. A weakness of the Bonferroni procedure is that it often overcorrects, making testing results too conservative because of a decrease in statistical power.

A variety of other procedures have been developed to control the overall Type I error level when multiple tests are performed. Some of these other multiple comparison and multiple testing procedures, including the Student–Newman–Keuls procedure, are derivatives of the Bonferroni procedure, modified to make the procedure less conservative without sacrificing Type I error control. Other multiple comparison and multiple testing procedures are simulation based and are not directly related to the Bonferroni procedure.

This entry describes the procedure's background, explains the procedure, and provides an example. This entry also presents applications for the procedure and examines recent research.

## Background

The Bonferroni procedure is named after the Italian mathematician Carlo Emilio Bonferroni. Although his work was in mathematical probability, researchers have since applied his work to statistical inference. Bonferroni's principal contribution to statistical inference was the identification of the probability inequality that bears his name.

## Explanation

The Bonferroni procedure is an application of the *Bonferroni inequality* to the probabilities associated with multiple testing. It prescribes using an adjustment to the significance level for individual tests when simultaneous statistical inference for several tests is being performed. The adjustment can be used for bounding simultaneous confidence intervals, as well as for simultaneous testing of hypotheses.

The Bonferroni inequality states the following:

1. Let $A_i, i = 1$ to $k$, represent $k$ events. Then, $P\left(\bigcap_{i=1}^{k} A_i\right) \geq 1 - \sum_{i=1}^{k} P(\overline{A}_i)$, where $\overline{A}_i$ is the complement of the event $Ai$.

2. Consider the mechanics of the Bonferroni inequality, $P\left(\bigcap_{i=1}^{k} A_i\right) \geq 1 - \sum_{i=1}^{k} P(\overline{A}_i)$,

3. and rewrite the inequality as follows:

$$1 - P\left(\bigcap_{i=1}^{k} A_i\right) \leq \sum_{i=1}^{k} P(\overline{A}_i).$$

Now, consider $\overline{A}_i$ as a Type I error in the $i$th test in a collection of $k$ hypothesis tests. Then $P\left(\bigcap_{i=1}^{k} A_i\right)$ represents the probability that no Type I errors occur in the $k$ hypothesis tests, and $1 - P\left(\bigcap_{i=1}^{k} A_i\right)$ represents the probability that at least one Type I error occurs in the $k$ hypothesis tests. $P(\overline{A}_i)$ represents the probability of a Type I error in the $i$th test, and we can label this probability as $\alpha_i = P(\overline{A}_i)$. So Bonferroni's inequality implies that the probability of at least one Type I error occurring in $k$ hypothesis tests is $\leq \sum_{i=1}^{k} \alpha_i$.

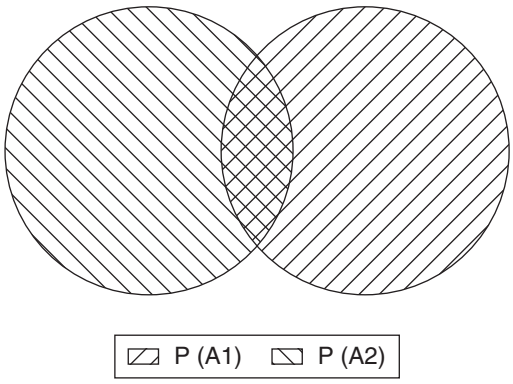



**Figure 1** Illustration of Bonferroni's Inequality

*Notes:* A1 = event 1; A2 = event 2; P(A1) = probability of A1; P(A2) = probability of A2. Interaction between events results in redundance in probability.

If, as is often assumed, all $k$ tests have the same probability of a Type I error, $\alpha$, then we can conclude that the probability of at least one Type I error occurring in $k$ hypothesis tests is $\leq k\alpha$.

Consider an illustration of Bonferroni's inequality in the simple case in which $k = 2$: Let the two events A1 and A2 have probabilities P(A1) and P(A2), respectively. The sum of the probabilities of the two events is clearly greater than the probability of the union of the two events because the former counts the probability of the intersection of the two events twice, as shown in Figure 1.

The Bonferroni procedure is simple in the sense that a researcher need only know the number of tests to be performed and the probability of a Type I error for those tests in order to construct this upper bound on the experiment-wise error rate. However, as mentioned earlier, the Bonferroni procedure is often criticized for being too conservative. Consider that the researcher does not typically know what the actual Type I error rate is for a given test. Rather, the researcher constructs the test so that the maximum allowable Type I error rate is $\alpha$. Then the actual Type I error rate may be considerably less than $\alpha$ for any given test.

For example, suppose a test is constructed with a nominal $\alpha = .05$. Suppose the researcher conducts $k = 10$ such tests on a given set of data, and the actual Type I error rate for each of the tests is .04. Using the Bonferroni procedure, the researcher concludes that the experiment-wise error rate is at most $10 * .05$, or .50. The error rate in this scenario is in fact .40, which is considerably less than .50.

As another example, consider the extreme case in which all $k = 10$ hypothesis tests are exactly dependent on each other—the same test is conducted 10 times on the same data. In this scenario, the experiment-wise error rate does not increase because of the multiple tests. In fact, if the Type I error rate for one of the tests is $\alpha = .05$, the experiment-wise error rate is the same, .05, for all 10 tests simultaneously. We can see this result from the Bonferroni inequality: $P\left(\bigcap_{i=1}^{k} A_i\right) = P(A_i)$ when the events, $A_i$, are all the same because $\bigcap_{i=1}^{k} A_i = A_i$. The Bonferroni procedure would suggest an upper bound on this experiment-wise probability as .50—overly conservative by 10-fold! It would be unusual for a researcher to conduct $k$ equivalent tests on the same data. However, it would not be unusual for a researcher to conduct $k$ tests and for many of those tests, if not all, to be partially interdependent. The more interdependent the tests are, the smaller the experiment-wise error rate and the more overly conservative the Bonferroni procedure is.

Other procedures have sought to correct for inflation in experiment-wise error rates without being as conservative as the Bonferroni procedure. However, none are as simple to use. These other procedures include the Student–Newman–Keuls, Tukey, and Scheffé procedures, to name a few. Descriptions of these other procedures and their uses can be found in many basic statistical methods textbooks, as well as this encyclopedia.

## Example

Consider the case of a researcher studying the effect of three different teaching methods on the average words per minute $(\mu_1, \mu_2, \mu_3)$ at which a student can read. The researcher tests three hypotheses: $\mu_1 = \mu_2 (\text{vs. } \mu_1 \neq \mu_2), \mu_1 = \mu_3 (\text{vs. } \mu_1 \neq \mu_3)$, and $\mu_2 = \mu_3 (\text{vs. } \mu_2 \neq \mu_3)$. Each test is conducted at a nominal level, $\alpha_0 = .05$, resulting in a comparison-wise error rate of $\alpha_c = .05$ for each test. Denote A1,

A2, and A3 as the event of falsely rejecting the null hypotheses 1, 2, and 3, respectively, and denote $p1, p2$, and $p3$ the probability of events A1, A2, and A3, respectively. These would be the individual $p$ values for these tests. It may be assumed that some dependence exists among the three events, A1, A2, and A3, principally because the events are all based on data collected from a single study. Consequently, the experiment-wise error rate, the probability of falsely rejecting any of the three null hypotheses, is at least equal to $\alpha_e = .05$ but potentially as large as $.05 * 3 = .15$. For this reason, we may apply the Bonferroni procedure by dividing our nominal level of $\alpha_0 = .05$ by $k = 3$ to obtain $\alpha_0^* = .0167$. Then, rather than comparing the $p$ values $p1$, $p2$, and $p3$ to $\alpha_0 = 05$, we compare them to $\alpha_0^* = .0167$. The experiment-wise error rate is therefore adjusted down so that it is less than or equal to the original intended nominal level of $\alpha_0 = .05$.

It should be noted that although the Bonferroni procedure is often used in the comparison of multiple means, because the adjustment is made to the nominal level, $\alpha_0$, or to the test's resulting $p$ value, the multiple tests could be hypothesis tests of any population parameters based on any probability distributions. So, for example, one experiment could involve a hypothesis test regarding a mean and another hypothesis test regarding a variance, and an adjustment based on $k = 2$ could be made to the two tests to maintain the experiment-wise error rate at the nominal level.

## Applications

As noted above, the Bonferroni procedure is used primarily to control the overall $\alpha$ level (i.e., the experiment-wise level) when multiple tests are being performed. Many statistical procedures have been developed at least partially for this purpose; however, most of those procedures have applications exclusively in the context of making multiple comparisons of group means after finding a significant ANOVA result. While the Bonferroni procedure can also be used in this context, one of its advantages over other such procedures is that it can also be used in other multiple testing situations that do not initially entail an omnibus test such as ANOVA.

For example, although most statistical tests do not advocate using a Bonferroni adjustment when testing beta coefficients in a multiple regression analysis, it has been shown that the overall Type I error rate in such an analysis involving as few as eight regression coefficients can exceed .30, resulting in almost a 1 in 3 chance of falsely rejecting a null hypothesis. Using a Bonferroni adjustment when one is conducting these tests would control that overall Type I error rate. Similar adjustments can be used to test for main effects and interactions in ANOVA and multivariate ANOVA designs because all that is required to make the adjustment is that the researcher knows the number of tests being performed. The Bonferroni adjustment has been used to adjust the experiment-wise Type I error rate for multiple tests in a variety of disciplines, such as medical, educational, and psychological research, to name a few.

## Recent Research

One of the main criticisms of the Bonferroni procedure is the fact that it overcorrects the overall Type I error rate, which results in lower statistical power. Many modifications to this procedure have been proposed over the years to try to alleviate this problem. Most of these proposed alternatives can be classified either as *step-down procedures* (e.g., the Holm method), which test the most significant (and, therefore, smallest) $p$ value first, or *step-up procedures* (e.g., the Hochberg method), which begin testing with the least significant (and largest) $p$ value. With each of these procedures, although the tests are all being conducted concurrently, each hypothesis is not tested at the same time or at the same level of significance.

More recent research has attempted to find a divisor between 1 and $k$ that would protect the overall Type I error rate at or below the nominal .05 level but closer to that nominal level so as to have a lesser effect on the power to detect actual differences. This attempt was based on the premise that making no adjustment to the $\alpha$ level is too liberal an approach (inflating the experiment-wise error rate), and dividing by the number of tests, $k$, is too conservative (overadjusting that error rate). It was shown that the optimal divisor is directly determined by the proportion of nonsignificant differences or relationships in the multiple tests being performed. Based on this result, a divisor of

$k(1 - q)$, where $q =$ the proportion of nonsignificant tests, did the best job of protecting against Type I errors without sacrificing as much power. Unfortunately, researchers often do not know, a priori, the number of nonsignificant tests that will occur in the collection of tests being performed. Consequently, research has also shown that a practical choice of the divisor is $k/1.5$ (rounded to the nearest integer) when the number of tests is greater than three. This modified Bonferroni adjustment will outperform alternatives in keeping the experiment-wise error rate at or below the nominal .05 level and will have higher power than other commonly used adjustments.

*Jamis J. Perrett and Daniel J. Mundfrom*

***See also*** Analysis of Variance (ANOVA); Hypothesis; Multiple Comparison Tests; Newman–Keuls Test and Tukey Test; Scheffé Test

**Further Reqdings**

Bain, L. J., & Engelhandt, M. (1992). *Introduction to probability and mathematical statistics* (2nd ed.). Boston: PWS-Kent. .

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika, 75*(4), 800–802.

Holland, B. S., & Copenhaver, M. D. (1987). An improved sequentially rejective Bonferroni test procedure. *Biometrics, 43,* 417–423.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6,* 65–70.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika, 75*(2), 383–386.

Mundfrom, D. J., Perrett, J., Schaffer, J., Piccone, A., & Roozeboom, M. A. (2006). Bonferroni adjustments in tests for regression coefficients. *Multiple Linear Regression Viewpoints, 32*(1), 1–6.

Rom, D. M. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality. *Biometrika, 77*(3), 663–665.

Roozeboom, M. A., Mundfrom, D. J., & Perrett, J. (2008, August). *A single-step modified Bonferroni procedure for multiple tests.* Paper presented at the Joint Statistical Meetings, Denver, CO.

Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association, 81*(395), 826–831.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika, 73*(3), 751–754.

Westfall, P. H., Tobias, R. D., Rom, D., Wolfinger, R. D., & Hochberg, Y. (1999). *Multiple comparisons and multiple tests using SAS.* Cary, NC: SAS Institute.

# BOOTSTRAPPING

The bootstrap is a computer-based statistical technique that is used to obtain measures of precision of parameter estimates. Although the technique is sufficiently general to be used in time-series analysis, permutation tests, cross-validation, nonlinear regression, and cluster analysis, its most common use is to compute standard errors and confidence intervals. Introduced by Bradley Efron in 1979, the procedure itself belongs in a broader class of estimators that use sampling techniques to create empirical distributions by resampling from the original data set. The goal of the procedure is to produce analytic expressions for estimators that are difficult to calculate mathematically. The name itself derives from the popular story in which Baron von Munchausen (after whom Munchausen syndrome is also named) was stuck at the bottom of a lake with no alternative but to grab his own bootstraps and pull himself to the surface. In a similar sense, when a closed-form mathematical solution is not easy to calculate, the researcher has no alternative but to "pull himself or herself up by the bootstraps" by employing such resampling techniques. This entry explores the basic principles and procedures of bootstrapping and examines its other applications and limitations.

## Basic Principles and Estimation Procedures

The fundamental principle on which the procedure is based is the belief that under certain general conditions, the relationship between a bootstrapped estimator and a parameter estimate should be similar to the relationship between the parameter estimate and the unknown population parameter of interest. As a means of better understanding the origins of this belief, Peter Hall suggested a valuable visual: a nested Russian doll. According to Hall's thought experiment, a researcher is interested in

determining the number of freckles present on the outermost doll. However, the researcher is not able to directly observe the outermost doll and instead can only directly observe the inner dolls, all of which resemble the outer doll, but because of their successively smaller size, each possesses successively fewer freckles. The question facing the researcher then is how to best use information from the observable inner dolls to draw conclusions about the likely number of freckles present on the outermost doll. To see how this works, assume for simplicity that the Russian doll set consists of three parts, the outermost doll and two inner dolls. In this case, the outermost doll can be thought of as the population, which is assumed to possess $n_0$ freckles; the second doll can be thought of as the original sample, which is assumed to possess $n_1$ freckles; and the third doll can be thought of as the bootstrap sample, which is assumed to possess $n_2$ freckles. A first guess in this situation might be to use the observed number of freckles on the second doll as the best estimate of the likely number of freckles on the outermost doll. Such an estimator will necessarily be biased, however, because the second doll is smaller than the outermost doll and necessarily possesses a smaller number of freckles. In other words, employing $n_1$ as an estimate of $n_0$ necessarily underestimates the true number of freckles on the outermost doll. This is where the bootstrapped estimator, $n_2$, reveals its true value. Because the third doll is smaller than the second doll by an amount similar to that by which the second doll is smaller than the outermost doll, the ratio of the number of freckles on the two inner dolls, $n_1 : n_2$, should be a close approximation of the ratio of the number of freckles on the second doll to number on the outer doll, $n_0 : n_1$. This in a nutshell is the principle underlying the bootstrap procedure.

More formally, the nonparametric bootstrap derives from an empirical distribution function,$\hat{F}$, which is a random sample of size $n$ from a probability distribution $F$. The estimator, $\hat{\theta}$, of the population parameter $\theta$ is defined as some function of the random sample $(X_1, X_2, \ldots X_n)$. The objective of the bootstrap is to assess the accuracy of the estimator, $\hat{\theta}$. The bootstrap principle described above states that the relationship between $\hat{\theta}$ and $\theta$ should be mimicked by that between $\theta^b$ and $\hat{\theta}$, where $\theta^b$ is the bootstrap estimator from bootstrap samples. In practice, bootstrap samples are obtained by a Monte Carlo procedure to draw (with replacement) multiple random samples of size $n$ from the initial sample data set, calculating the parameter of interest for the sample drawn, say $\theta^b$, and repeating the process $k$ times. Hence, the bootstrap technique allows researchers to generate an estimated sampling distribution in cases in which they have access to only a single sample rather than the entire population. A minimum value for $k$ is typically assumed to be 100 and can be as many as 10,000, depending on the application.

Peter Bickel and David Freedman defined the following three necessary conditions if the bootstrap is to provide consistent estimates of the asymptotic distribution of a parameter: (1) The statistic being bootstrapped must converge weakly to an asymptotic distribution whenever the data-generating distribution is in a neighborhood of the truth, or in other words, the convergence still occurs if the truth is allowed to change within the neighborhood as the sample size grows. (2) The convergence to the asymptotic distribution must be uniform in that neighborhood. (3) The asymptotic distribution must depend on the data-generating process in a continuous way. If all three conditions hold, then the bootstrap should provide reliable estimates in many different applications.

As a concrete example, assume that we wish to obtain the standard error of the median value for a sample of 30 incomes. The researcher needs to create 100 bootstrap samples because this is the generally agreed on number of replications needed to compute a standard error. The easiest way to sample with replacement is to take the one data set and copy it 500 times for 100 bootstrap samples in order to guarantee that each observation has an equal likelihood of being chosen in each bootstrap sample. The researcher then assigns random numbers to each of the 15,000 observations $(500 * 30)$ and sorts each observation by its random number assignment from lowest to highest. The next step is to make 100 bootstrap samples of 30 observations each and disregard the other 12,000 observations. After the 100 bootstrap samples have been made, the median is calculated from each of the samples, and the bootstrap estimate of the standard error is just the standard deviation of the 100 bootstrapped medians. Although this procedure may seem complicated, it

is actually relatively easy to write a bootstrapping program with the use of almost any modern statistical program, and in fact, many statistical programs include a bootstrap command.

Besides generating standard error estimates, the bootstrap is commonly used to directly estimate confidence intervals in cases in which they would otherwise be difficult to produce. Although a number of different bootstrapping approaches exist for computing confidence intervals, the following discussion focuses on two of the most popular. The first, called *the percentile method,* is straightforward and easy to implement. For illustration purposes, assume that the researcher wishes to obtain a 90% confidence interval. To do so, the researcher would (a) start by obtaining 1,000 bootstrap samples and the resulting 1,000 bootstrap estimates, $\hat{\theta}^b$, and (b) order the 1,000 observed estimates from the smallest to the largest. The 90% confidence interval would then consist of the specific value bootstrap estimates falling at the 5th and the 95th percentiles of the sorted distribution. This method typically works well for large sample sizes because the bootstrap mimics the sampling distribution, but it does not work well for small sample size. If the number of observations in the sample is small, Bradley Efron and Robert Tibshirani have suggested using a bias correction factor.

The second approach, called the *bootstrap* t *confidence interval,* is more complicated than the percentile method, but it is also more accurate. To understand this method, it is useful to review a standard confidence interval, which is defined as $[\hat{\theta} - t_{\alpha/2,df} s.e.(\hat{\theta}), \hat{\theta} + t_{\alpha/2,df} s.e.(\hat{\theta})]$, where $\hat{\theta}$ is the estimate, $t_{\alpha/2,df}$ is the critical value from the *t*-table with *df* degrees of freedom for a $(1 - \alpha)$ confidence interval, and $s.e.(\hat{\theta})$ is the standard error of the estimate. The idea behind the bootstrap *t* interval is that the critical value is found through bootstrapping instead of simply reading the value contained in a published table. Specifically, the bootstrap *t* is defined as $T^{boot} = (\hat{\theta}^{boot} - \hat{\theta})/S^{boot}$, where $\hat{\theta}^{boot}$ is the estimate of $\theta$ from a bootstrap sample and $S^{boot}$ is an estimate of the standard deviation of $\theta$ from the bootstrap sample. The $k$ values of $T^{boot}$ are then ordered from lowest to highest, and then, for a 90% confidence interval, the value at the 5th percentile is the lower critical value and the value at the 95th percentile is the higher critical value. Thus the bootstrapped *t* interval is

$$[\hat{\theta} - T^{boot}_{.95}\ s.e.(\hat{\theta}), \hat{\theta} - T^{boot}_{.05} s.e.(\hat{\theta})].$$

Michael Chernick has pointed out that the biggest drawback of this method is that it is not always obvious how to compute the standard errors, $S^{boot}$ and $s.e.(\hat{\theta})$.

## Other Applications

In addition to calculating such measures of precision, the bootstrap procedure has gained favor for a number of other applications. For one, the bootstrap is now popular as a method for performing bias reduction. Bias reduction can be explained as follows. The bias of an estimator is the difference between the expected value of an estimator, $E(\hat{\theta})$, and the true value of the parameter, $\theta$, or $E(\hat{\theta} - \theta)$. If an estimator is biased, then this value is nonzero, and the estimator is wrong on average. In the case of such a biased estimator, the bootstrap principle is employed such that the bias is estimated by taking the average of the difference between the bootstrap estimate, $\hat{\theta}^b$, and the estimate from the initial sample, $\hat{\theta}$ over the $k$ different bootstrap estimates. Efron defined the bias of the bootstrap as $E(\hat{\theta} - \hat{\theta}^b)$ and suggested reducing the bias of the original estimator, $\hat{\theta}$, by adding estimated bias. This technique produces an estimator that is close to unbiased.

Recently, the bootstrap has also become popular in different types of regression analysis, including linear regression, nonlinear regression, time-series analysis, and forecasting. With linear regression, the researcher can either fit the residuals from the fitted model, or the vector of the dependent and independent variables can be bootstrapped. If the error terms are not normal and the sample size is small, then the researcher is able to obtain bootstrapped confidence intervals, like the one described above, instead of relying on asymptotic theory that likely does not apply. In nonlinear regression analysis, the bootstrap is a very useful tool because there is no need to differentiate and an analytic expression is not necessary.

### Limitations

Although the above discussion has highlighted that the bootstrap technique is potentially valuable in a number of situations, it should be noted that it is not the ideal solution to every statistical problem. One problem would occur in cases in which parameters are constrained to be on a boundary of the parameter space (such as when a priori theoretical restrictions require a certain estimated parameter to be of a specific sign). Common examples of such restrictions include traditional demand analysis in which the income effect for a normal good is constrained to be positive and the own-price effect is constrained to be negative, cost function analysis in which curvature constraints imply that second-order price terms satisfy concavity conditions, and time-series models for conditional heteroskedasticity in which the same parameters are constrained to be nonnegative. Such cases are potentially problematic for the researcher because standard error estimates and confidence bounds are difficult to compute using classical statistical inference, and therefore the bootstrap would be a natural choice. Don Andrews has demonstrated, however, that this procedure is not asymptotically correct to the first order when parameters are on a boundary. This is because the bootstrap puts too much mass below the cutoff point for the parameter and therefore does a poor job of mimicking the true population distribution. Other circumstances in which the bootstrap fails include an extremely small sample size, its use with matching estimators to evaluate programs, and distributions with long tails.

*Christiana Hilmer*

***See also*** Bias; Confidence Intervals; Distribution; Jackknife; Central Tendency, Measures of; Median; Random Sampling; Sampling; Standard Deviation; Standard Error of Estimate; Statistic; Student's *t* Test; Unbiased Estimator; Variability, Measure of

### Further Readings

Abadie, A., & Imbens, G. W. (2006). *On the failure of the bootstrap for matching estimators* (NBER Working Paper 0325). Cambridge, MA: National Bureau of Economic Research.

Andrews, D. W. K. (1999). Estimation when a parameter is on a boundary. *Econometrica, 67*(6), 1341–1383.

Andrews, D. W. K. (2003). A three-step method for choosing the number of bootstrap repetitions. *Econometrica, 68*(1), 23–51.

Arthreya, K. B. (1985). Bootstrap of the mean in the infinite variance case. *Annals of Statistics, 15,* 724–731.

Bickel, P. J., & Freeman, D. (1981). Some asymptotic theory for the bootstrap. *Annals of Statistics, 9,* 1196–1217.

Chernick, M. R. (1999). *Bootstrap methods: A practitioner's guide*. New York: Wiley.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics, 7,* 1–26.

Efron, B. (1983). Estimating the error rate of a perdiction rule: Improvements on cross-validation. *Journal of the American Statistical Association, 78,* 316–331.

Efron, B. (1987). Better bootstrap confidence intervals, *Journal of the American Statistical Association, 82,* 171–200.

Efron, B., & Tibshirani, R. (1986). *An introduction to the bootstrap*. New York: Chapman & Hall.

Hall, P. (1992). *The bootstrap and Edgeworth expansion.* New York: Springer-Verlag.

## Box-and-Whisker Plot

A box-and-whisker plot, or box plot, is a tool used to visually display the range, distribution symmetry, and central tendency of a distribution in order to illustrate the variability and the concentration of values within a distribution. The box plot is a graphical representation of the five-number summary, or a quick way of summarizing the center and dispersion of data for a variable. The five-number summary includes the minimum value, 1st (lower) quartile (Q1), median, 3rd (upper) quartile (Q3), and the maximum value. Outliers are also indicated on a box plot. Box plots are especially useful in research methodology and data analysis as one of the many ways to visually represent data. From this visual representation, researchers glean several pieces of information that may aid in drawing conclusions, exploring unexpected patterns in the data, or prompting the researcher to develop future research questions and hypotheses. This entry provides an overview of the history of the box plot, key components and construction of the

box plot, and a discussion of the appropriate uses of a box plot.

## History

A box plot is one example of a graphical technique used within exploratory data analysis (EDA). EDA is a statistical method used to explore and understand data from several angles in social science research. EDA grew out of work by John Tukey and his associates in the 1960s and was developed to broadly understand the data, graphically represent data, generate hypotheses and build models to guide research, add robust measures to an analysis, and aid the researcher in finding the most appropriate method for analysis. EDA is especially helpful when the researcher is interested in identifying any unexpected or misleading patterns in the data. Although there are many forms of EDA, researchers must employ the most appropriate form given the specific procedure's purpose and use.

## Definition and Construction

One of the first steps in any statistical analysis is to describe the central tendency and the variability of the values for each variable included in the analysis. The researcher seeks to understand the center of the distribution of values for a given variable (*central tendency*) and how the rest of the values fall in relation to the center (*variability*). Box plots are used to visually display variable distributions through the display of robust statistics, or statistics that are more resistant to the presence of outliers in the data set. Although there are somewhat different ways to construct box plots depending on the way in which the researcher wants to display outliers, a box plot always provides a visual display of the five-number summary. The median is defined as the value that falls in the middle after the values for the selected variable are ordered from lowest to highest value, and it is represented as a line in the middle of the rectangle within a box plot. As it is the central value, 50% of the data lie above the median and 50% lie below the median. When the distribution contains an odd number of values, the median represents an actual value in the distribution. When the distribution contains an even number of values, the median represents an average of the two middle values.

To create the rectangle (or box) associated with a box plot, one must determine the 1st and 3rd quartiles, which represent values (along with the median) that divide all the values into four sections, each including approximately 25% of the values. The 1st (lower) quartile ($Q_1$) represents a value that divides the lower 50% of the values (those below the median) into two equal sections, and the 3rd (upper) quartile ($Q_3$) represents a value that divides the upper 50% of the values (those above the median) into two equal sections. As with calculating the median, quartiles may represent the average of two values when the number of values below and above the median is even. The rectangle of a box plot is drawn such that it extends from the 1st quartile through the 3rd quartile and thereby represents the *interquartile range* (IQR; the distance between the 1st and 3rd quartiles). The rectangle includes the median.

In order to draw the "whiskers" (i.e., lines extending from the box), one must identify *fences,* or values that represent minimum and maximum values that would not be considered outliers. Typically, fences are calculated to be $Q - 1.5$ IQR (lower fence) and $Q_3 + 1.5$ IQR (upper fence). Whiskers are lines drawn by connecting the most extreme values that fall within the fence to the lines representing $Q_1$ and $Q_3$. Any value that is greater than the upper fence or lower than the lower fence is considered an outlier and is displayed as a special symbol beyond the whiskers. Outliers that extend beyond the fences are typically considered *mild outliers* on the box plot. An *extreme outlier* (i.e., one that is located beyond 3 times the length of the IQR from the 1st quartile (if a low outlier) or 3rd quartile (if a high outlier) may be indicated by a different symbol. Figure 1 provides an illustration of a box plot.

Box plots can be created in either a vertical or a horizontal direction. (In this entry, a vertical box plot is generally assumed for consistency.) They can often be very helpful when one is attempting to compare the distributions of two or more data sets or variables on the same scale, in which case they can be constructed side by side to facilitate comparison.

| *Data set values* | *Defining features of this box plot* |
|---|---|
| 2.0 | Median = 6.0 |
| 2.0 | First (Lower) Quartile = 3.0 |
| 2.0 | Third (Upper) Quartile = 8.0 |
| 3.0 | Interquartile Range (IQR) = 5.0 |
| 3.0 | Lower Inner Fence = -4.5 |
| 5.0 | Upper Inner Fence = 15.5 |
| 6.0 | Range = 20.0 |
| 6.0 | Mild Outlier = 22.0 |
| 7.0 | |
| 7.0 | |
| 8.0 | |
| 8.0 | |
| 9.0 | |
| 10.0 | |
| 22.0 | |

**Figure 1** Box Plot Created With a Data Set and SPSS (an IBM company, formerly called PASW® Statistics)

*Notes:* Data set values: 2.0, 2.0, 2.0, 3.0, 3.0, 5.0, 6.0, 6.0, 7.0, 7.0, 8.0, 8.0, 9.0, 10.0, 22.0. Defining features of this box plot: Median = 6.0; First (lower) quartile = 3.0; Third (upper) quartile = 8.0; Interquartile range (IQR) = 5.0; Lower inner fence = 4.5; Upper inner fence = 15.5; Range = 20.0; Mild outlier = 22.0.

### *Steps to Creating a Box Plot*

The following six steps are used to create a vertical box plot:

1. Order the values within the data set from smallest to largest and calculate the median, lower quartile ($Q_1$), upper quartile ($Q_3$), and minimum and maximum values.
2. Calculate the IQR.
3. Determine the lower and upper fences.
4. Using a number line or graph, draw a box to mark the location of the 1st and

3rd quartiles. Draw a line across the box to mark the median.

5. Make a short horizontal line below and above the box to locate the minimum and maximum values that fall within the lower and upper fences. Draw a line connecting each short horizontal line to the box. These are the box plot whiskers.
6. Mark each outlier with an asterisk or an "o."

### *Making Inferences*

R. Lyman Ott and Michael Longnecker described five inferences that one can make from a box plot. First, the researcher can easily identify the median of the data by locating the line drawn in the middle of the box. Second, the researcher can easily identify the variability of the data by looking at the length of the box. Longer boxes illustrate greater variability whereas shorter box lengths illustrate a tighter distribution of the data around the median. Third, the researcher can easily examine the symmetry of the middle 50% of the data distribution by looking at where the median line falls in the box. If the median is in the middle of the box, then the data are evenly distributed on either side of the median, and the distribution can be considered symmetrical. Fourth, the researcher can easily identify outliers in the data by the asterisks outside the whiskers. Fifth, the researcher can easily identify the skewness of the distribution. On a distribution curve, data skewed to the right show more of the data to the left with a long "tail" trailing to the right. The opposite is shown when the data are skewed to the left. To identify skewness on a box plot, the researcher looks at the length of each half of the box plot. If the lower or left half of the box plot appears longer than the upper or right half, then the data are skewed in the lower direction or skewed to the left. If the upper half of the box plot appears longer than the lower half, then the data are skewed in the upper direction or skewed to the right. If a researcher suspects the data are skewed, it is recommended that the researcher investigate further by means of a histogram.

## Variations

Over the past few decades, the availability of several statistical software packages has made EDA easier for social science researchers. However, these statistical packages may not calculate parts of a box plot in the same way, and hence some caution is warranted in their use. One study conducted by Michael Frigge, David C. Hoaglin, and Boris Iglewicz found that statistical packages calculate aspects of the box plot in different ways. In one example, the authors used three different statistical packages to create a box plot with the same distribution. Though the median looked approximately the same across the three box plots, the differences appeared in the length of the whiskers. The reason for the differences was the way the statistical packages used the interquartile range to calculate the whiskers. In general, to calculate the whiskers, one multiplies the interquartile range by a constant and then adds the result to $Q_3$ and subtracts it from $Q_1$. Each package used a different constant, ranging from 1.0 to 3.0. Though packages typically allow the user to adjust the constant, a package typically sets a default, which may not be the same as another package's default. This issue, identified by Frigge and colleagues, is important to consider because it guides the identification of outliers in the data. In addition, such variations in calculation lead to the lack of a standardized process and possibly to consumer confusion. Therefore, the authors provided three suggestions to guide the researcher in using statistical packages to create box plots. First, they suggested using a constant of 1.5 when the number of observations is between 5 and 20. Second, they suggested using a constant of 2.0 for outlier detection and rejection. Finally, they suggested using a constant of 3.0 for extreme cases. In the absence of standardization across statistical packages, researchers should understand how a package calculates whiskers and follow the suggested constant values.

## Applications

As with all forms of data analysis, there are many advantages and disadvantages, appropriate uses, and certain precautions researchers should consider

when using a box plot to display distributions. Box plots provide a good visualization of the range and potential skewness of the data. A box plot may provide the first step in exploring unexpected patterns in the distribution because box plots provide a good indication of how the data are distributed around the median. Box plots also clearly mark the location of mild and extreme outliers in the distribution. Other forms of graphical representation that graph individual values, such as dot plots, may not make this clear distinction. When used appropriately, box plots are useful in comparing more than one sample distribution side by side. In other forms of data analysis, a researcher may choose to compare data sets using a *t* test to compare means or an *F* test to compare variances. However, these methods are more vulnerable to skewness in the presence of extreme values. These methods must also meet normality and equal variance assumptions. Alternatively, box plots can compare the differences between variable distributions without the need to meet certain statistical assumptions.

However, unlike other forms of EDA, box plots show less detail than a researcher may need. For one, box plots may display only the five-number summary. They do not provide frequency measures or the quantitative measure of variance and standard deviation. Second, box plots are not used in a way that allows the researcher to compare the data with a normal distribution, which stem plots and histograms do allow. Finally, box plots would not be appropriate to use with a small sample size because of the difficulty in detecting outliers and finding patterns in the distribution.

Besides taking into account the advantages and disadvantages of using a box plot, one should consider a few precautions. In a 1990 study conducted by John T. Behrens and colleagues, participants frequently made judgment errors in determining the length of the box or whiskers of a box plot. In part of the study, participants were asked to judge the length of the box by using the whisker as a judgment standard. When the whisker length was longer than the box length, the participants tended to overestimate the length of the box. When the whisker length was shorter than the box length, the participants tended to underestimate the length of the box. The same result was found when the participants judged the length of the whisker by using the box length as a judgment standard. The study also found that compared with vertical box plots, box plots positioned horizontally were associated with fewer judgment errors.

*Sara C. Lewandowski and Sara E. Bolt*

***See also*** Exploratory Data Analysis; Histogram; Outlier

**Further Readings**

Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. *Psychological Methods, 2,* 131–160.

Behrens, J. T., Stock, W. A., & Sedgwick, C. E. (1990). Judgment errors in elementary box-plot displays. *Communications in Statistics B: Simulation and Computation, 19,* 245–262.

Frigge, M., Hoaglin, D. C., & Iglewicz, B. (1989). Some implementations of the boxplot. *American Statistician, 43,* 50–54.

Massart, D. L., Smeyers-Verbeke, J., Capron, X., & Schlesier, K. (2005). Visual presentation of data by means of box plots. *LCGC Europe, 18,* 215–218.

Moore, D. S. (2001). *Statistics: Concepts and controversies* (5th ed.). New York: W. H. Freeman.

Moore, D. S., & McCabe, P. G. (1998). *Introduction to the practice of statistics* (3rd ed.). New York: W. H. Freeman.

Ott, R. L., & Longnecker, M. (2001). *An introduction to statistical methods and data analysis* (5th ed.). Pacific Grove, CA: Wadsworth.

Tukey, J. W. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley.

# b Parameter

The b parameter is an item response theory (IRT)–based index of item difficulty. As IRT models have become an increasingly common way of modeling item response data, the b parameter has become a popular way of characterizing the difficulty of an individual item, as well as comparing the relative difficulty levels of different items. This entry addresses the b parameter with regard to different IRT models. Further, it discusses interpreting, estimating, and studying the b parameter.

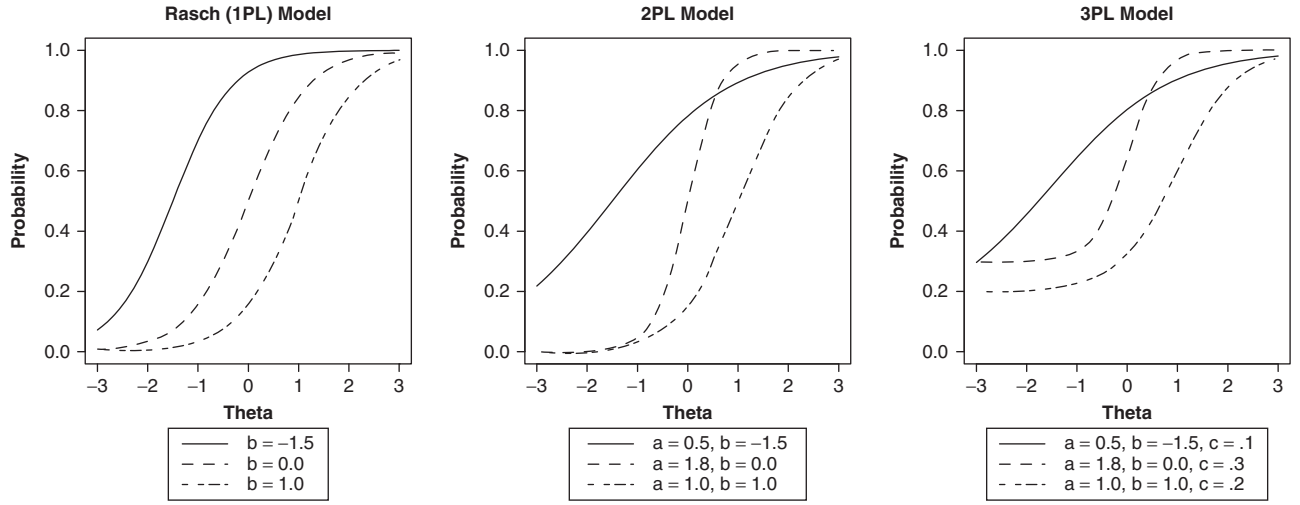



**Figure 1** Item Characteristic Curves for Example Items, One-Parameter Logistic (1PL or Rasch), Two-Parameter Logistic (2PL), and Three-Parameter Logistic (3PL) Models

## b Parameter Within Different Item Response Theory Models

The precise interpretation of the b parameter is dependent on the specific IRT model within which it is considered, the most common being the one-parameter logistic (1PL) or Rasch model, the two-parameter logistic (2PL) model, and three-parameter logistic (3PL) model. Under the 1PL model, the b parameter is the single item feature by which items are distinguished in characterizing the likelihood of a correct response. Specifically, the probability of correct response ($X_{ij} = 1$) by examinee $i$ to item $j$ is given by

$$P(X_{ij} = 1) = \frac{\exp(\theta_i - b_j)}{1 + \exp(\theta_i - b_j)},$$

where $\theta_j$ represents an ability-level (or trait-level) parameter of the examinee. An interpretation of the b parameter follows from its being attached to the same metric as that assigned to $\theta$.

Usually this metric is continuous and unbounded; the indeterminacy of the metric is often handled by assigning either the mean of $\theta$ (across examinees) or b (across items) to 0. Commonly b parameters will assume values between $-3$ and 3, with more extreme positive values representing more difficult (or infrequently endorsed) items, and more extreme negative values representing easy (or frequently endorsed) items.

The 2PL and 3PL models include additional item parameters that interact with the b parameter in determining the probability of correct response. The 2PL model adds an item discrimination parameter ($a_j$), so the probability of correct response is

$$P(X_{ij} = 1) = \frac{\exp[a_j(\theta_i - b_j)]}{1 + \exp[a_j(\theta_i - b_j)]},$$

and the 3PL model adds a lower asymptote ("pseudoguessing") parameter, resulting in

$$P(X_{ij} = 1) = c_j + (1 - c_j)\frac{\exp[a_j(\theta_i - b_j)]}{1 + \exp[a_j(\theta_i - b_j)]}.$$

While the same general interpretation of the b parameter as a difficulty parameter still applies under the 2PL and 3PL models, the discrimination and lower asymptote parameters also contribute to the likelihood of a correct response at a given ability level.

## Interpretation of the b Parameter

Figure 1 provides an illustration of the b parameter with respect to the 1PL, 2PL, and 3PL models. In

this figure, item characteristic curves (ICCs) for three example items are shown with respect to each model. Each curve represents the probability of a correct response as a function of the latent ability level of the examinee. Across all three models, it can be generally seen that as the b parameter increases, the ICC tends to decrease, implying a lower probability of correct response.

In the 1PL and 2PL models, the b parameter has the interpretation of representing the level of the ability or trait at which the respondent has a .50 probability of answering correctly (endorsing the item). For each of the models, the b parameter also identifies the ability level that corresponds to the inflection point of the ICC, and thus the b parameter can be viewed as determining the ability level at which the item is maximally informative. Consequently, the b parameter is a critical element in determining where along the ability continuum an item provides its most effective estimation of ability, and thus the parameter has a strong influence on how items are selected when administered adaptively, such as in a computerized adaptive testing environment.

Under the 1PL model, the b parameter effectively orders all items from easiest to hardest, and this ordering is the same regardless of the examinee ability or trait level. This property is no longer present in the 2PL and 3PL models, as the ICCs of items may cross, implying a different ordering of item difficulties at different ability levels. This property can also be seen in the example items in Figure 1 in which the ICCs cross for the 2PL and 3PL models, but not for the 1PL model. Consequently, while the b parameter remains the key factor in influencing the difficulty of the item, it is not the sole determinant.

An appealing aspect of the b parameter for all IRT models is that its interpretation is invariant with respect to examinee ability or trait level. That is, its value provides a consistent indicator of item difficulty whether considered for a population of high, medium, or low ability. This property is not present in more classical measures of item difficulty (e.g., "proportion correct"), which are influenced not only by the difficulty of the item, but also by the distribution of ability in the population in which they are administered. This invariance property allows the b parameter to play a fundamental role in how important measurement applications, such as item bias (differential item functioning), test equating, and appropriateness measurement, are conducted and evaluated in an IRT framework.

## Estimating the b Parameter

The b parameter is often characterized as a structural parameter within an IRT model and as such will generally be estimated in the process of fitting an IRT model to item response data. Various estimation strategies have been proposed and investigated, some being more appropriate for certain model types. Under the 1PL model, conditional maximum likelihood procedures are common. For all three model types, marginal maximum likelihood, joint maximum likelihood, and Bayesian estimation procedures have been developed and are also commonly used.

## Studying the b Parameter

The b parameter can also be the focus of further analysis. Models such as the *linear logistic test model* and its variants attempt to relate the b parameter to task components within an item that account for its difficulty. Such models also provide a way in which the b parameter's estimates of items can ultimately be used to validate a test instrument. When the b parameter assumes the value expected given an item's known task components, the parameter provides evidence that the item is functioning as intended by the item writer.

*Daniel Bolt*

***See also*** Differential Item Functioning; Item Analysis; Item Response Theory; Parameters; Validity of Measurement

### Further Readings

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

# Canonical Correlation Analysis

Canonical correlation analysis (CCA) is a multivariate statistical method that analyzes the relationship between two *sets* of variables, in which each set contains at least two variables. It is the most general type of the general linear model, with multiple regression, multiple analysis of variance, analysis of variance, and discriminant function analysis all being special cases of CCA.

Although the method has been available for more than 70 years, its use has been somewhat limited until fairly recently due to its lack of inclusion in common statistical programs and its rather labor-intensive calculations. Currently, however, many computer programs do include CCA, and thus the method has become somewhat more widely used.

This entry begins by explaining the basic logic of and defining important terms associated with CCA. Next, this entry discusses the interpretation of CCA results, statistical assumptions, and limitations of CCA. Last, it provides an example from the literature.

## Basic Logic

The logic of CCA is fairly straightforward and can be explained best by likening it to a "multiple-multiple regression." That is, in multiple regression a researcher is interested in discovering the variables (among a set of variables) that best predict a single variable. The set of variables may be termed the *independent,* or *predictor, variables;* the single variable may be considered the *dependent,* or *criterion, variable.* CCA is similar, except that there are multiple dependent variables, as well as multiple independent variables. The goal is to discover the pattern of variables (on both sides of the equation) that combine to produce the highest predictive values for both sets. The resulting combination of variables for each side, then, may be thought of as a kind of latent or underlying variable that describes the relation between the two sets of variables.

A simple example from the literature illustrates its use: A researcher is interested in investigating the relationships among gender, social dominance orientation, right wing authoritarianism, and three forms of prejudice (stereotyping, opposition to equality, and negative affect). Gender, social dominance orientation, and right wing authoritarianism constitute the predictor set; the three forms of prejudice are the criterion set. Rather than computing three separate multiple regression analyses (viz., the three predictor variables regressing onto one criterion variable, one at a time), the researcher instead computes a CCA on the two sets of variables to discern the most important predictor(s) of the three forms of prejudice. In this example, the CCA revealed that social dominance orientation emerged as the overall most important dimension that underlies all three forms of prejudice.

## Important Terms

To appreciate the various terms associated with CCA, it is necessary to have a basic understanding of the analytic procedure itself. The first step in CCA involves collapsing each person's score for each variable, in the two variable sets, into a single composite, or "synthetic," variable. These synthetic variables are created such that the correlation between the two sets is maximal. This occurs by weighting each person's score and then summing the weighted scores for the respective variable sets. Pairs of linear synthetic variables created by this maximization process are called *canonical variates*. The bivariate correlation between the pairs of variates is the *canonical correlation* (sometimes called the *canonical function*). There will be two canonical variates produced for each canonical correlation, with one variate representing the predictor variables and the other representing the criterion variables. The total number of canonical variate pairs produced is equal to the number of variables in either the criterion or predictor set, whichever is smaller. Finally, squaring the canonical correlation coefficient yields the proportion of variance the pairs of canonical variates (not the original variables) linearly share.

## Interpretation of Results

Similar to other statistical methods, the first step in CCA is determining the statistical significance of the canonical correlation coefficients associated with the variates. Different tests of significance (e.g., Wilks's lambda, Bartlett's test) are generally reported in statistical computer programs, and only those canonical correlations that differ significantly from zero are subsequently interpreted. With respect to the previously mentioned example (social dominance and right wing authoritarianism), three canonical variates were produced (because the smallest variable set contained three variables), but in that case, only the first two were significant. Canonical correlations are reported in descending order of importance, and normally only the first one or two variates are significant. Further, even if a canonical correlation is significant, most researchers do not interpret those below .30 because the amount of variance shared (that is, the correlation squared) is so small that the canonical variate is of little practical significance.

A second method of interpretation involves evaluating the degree to which the individual variables load onto their respective canonical variates. Variables that have high loadings on a particular variate essentially have more in common with it and thus should be given more weight in interpreting its meaning. This is directly analogous to factor loading in factor analysis and indicates the degree of importance of the individual variables in the overall relation between the criterion and predictor variable sets. The commonly accepted convention is to interpret only those variables that achieve loadings of .40 or higher. The term *structure coefficient* (or *canonical factor loadings*) refers to the correlation between each variable and its respective canonical variate, and its interpretation is similar to the bivariate correlation coefficient. In addition, as with correlation coefficient interpretation, both the direction and the magnitude of the structure coefficient are taken into consideration. The structure coefficient squared represents the amount of variance a given variable accounts for in its own canonical variate, based on the set of variables to which it belongs. Structure coefficients, then, represent the extent to which the individual variables and their respective canonical variates are related, and in effect, their interpretation forms the crux of CCA.

## Statistical Assumptions

Reliable use of CCA requires multivariate normality, and the analysis itself proceeds on the assumption that all variables and linear combinations of variables approximate the normal distribution. However, with large sample sizes, CCA can be robust to this violation.

The technique also requires minimal measurement error (e.g., alpha coefficients of .80 or above) because low scale reliabilities attenuate correlation coefficients, which, ultimately, reduces the probability of detecting a significant canonical correlation. In addition, it is very useful to have an adequate range of scores for each variable because correlation coefficients can also be attenuated by truncated or restricted variance; thus, as with other correlation methods, sampling methods that increase score variability are highly recommended.

Outliers (that is, data points that are well "outside" a particular distribution of scores) can also significantly attenuate correlation, and their occurrence should be minimized or, if possible, eliminated.

For conventional CCA, linear relationships among variables are required, although CCA algorithms for nonlinear relationships are currently available. And, as with other correlation-based analyses, low multicollinearity among variables is assumed. Multicollinearity occurs when variables in a correlation matrix are highly correlated with each other, a condition that reflects too much redundancy among the variables. As with measurement error, high multicollinearity also reduces the magnitudes of correlation coefficients.

Stable (i.e., reliable) canonical correlations are more likely to be obtained if the sample size is large. It is generally recommended that adequate samples should range from 10 to 20 cases per variable. So, for example, if there are 10 variables, one should strive for a minimum of 100 cases (or participants). In general, the smaller the sample size, the more unstable the CCA.

## Limitation

The main limitation associated with CCA is that it is often difficult to interpret the meaning of the resulting canonical variates. As others have noted, a mathematical procedure that maximizes correlations may not necessarily yield a solution that is maximally interpretable. This is a serious limitation and may be the most important reason CCA is not used more often. Moreover, given that it is a descriptive technique, the problems with interpreting the meaning of the canonical correlation and associated variates are particularly troublesome. For example, suppose a medical sociologist found that low unemployment level, high educational level, and high crime rate (the predictor variables) are associated with good medical outcomes, low medical compliance, and high medical expenses (the criterion variables). What might this pattern mean? Perhaps people who are employed, educated, and live in high crime areas have expensive, but good, health care, although they do not comply with doctors' however, compared with other multivariate techniques, with CCA there appears to be greater difficulty in extracting the meaning (i.e., the latent variable) from the obtained results.

## Example in the Literature

In the following example, drawn from the adolescent psychopathology literature, a CCA was performed on two sets of variables, one (the predictor set) consisting of personality pattern scales of the Millon Adolescent Clinical Inventory, the other (the criterion set) consisting of various mental disorder scales from the Adolescent Psychopathology Scale. The goal of the study was to discover whether the two sets of variables were significantly related (i.e., would there exist a multivariate relationship?), and if so, how might these two sets be related (i.e., how might certain personality styles be related to certain types of mental disorders?).

The first step in interpretation of a CCA is to present the significant variates. In the above example, four significant canonical variates emerged; however, only the first two substantially contributed to the total amount of variance (together accounting for 86% of the total variance), and thus only those two were interpreted. Second, the structure coefficients (canonical loadings) for the two significant variates were described. Structure coefficients are presented in order of absolute magnitude and are always interpreted as a pair. By convention, only coefficients greater than .40 are interpreted. In this case, for the first canonical variate, the predictor set accounted for 48% ($Rc^2$) of the variance in the criterion set. Further examination of the canonical loadings for this variate showed that the predictor set was represented mostly by the Conformity subscale (factor loading of .92) and was related to lower levels of mental disorder symptoms. The structure coefficients for the second canonical variate were then interpreted in a similar manner.

The last, and often most difficult, step in CCA is to interpret the overall meaning of the analysis. Somewhat similar to factor analysis interpretation, latent variables are inferred from the pattern of the structure coefficients for the variates. A possible interpretation of the above example might be that an outgoing conformist personality style is predictive of overall better mental health for adolescents at risk for psychopathology.

**Summary**

The purpose of canonical correlation analysis, a multivariate statistical technique, is to analyze the relationships between two sets of variables, in which each set contains at least two variables. It is appropriate for use when researchers want to know whether the two sets are related and, if so, how they are related. CCA produces canonical variates based on linear combinations of measured variables. The variates are a kind of latent variable that relates one set of variables to the other set. CCA shares many of the statistical assumptions of other correlational techniques. The major limitation of CCA lies in its interpretability; that is, although significant variates can be derived mathematically, they may not necessarily be meaningful.

*Tracie L. Blumentritt*

***See also*** Bivariate Regression; Multiple Regression

**Further Readings**

Hotelling, H. (1935). The most predictable criterion. *Journal of Educational Psychology, 26,* 139–142.
Levine, M. S. (1977). *Canonical analysis and factor comparison.* Beverly Hills, CA: Sage.
Stevens, J. (1986). *Applied multivariate statistics for the social sciences.* Hillsdale, NJ: Lawrence Erlbaum.
Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics* (3rd ed.). New York: HarperCollins.
Thompson, B. (1984). *Canonical correlation analysis: Uses and interpretation.* Beverly Hills, CA: Sage.

# Case-Only Design

Analytical studies are designed to test the hypotheses created by descriptive studies and to assess the cause–effect association. These studies are able to measure the effect of a specific exposure on the occurrence of an outcome over time. Depending on the nature of the exposure, whether it has been caused experimentally as an intervention on the study subjects or has happened naturally, without any specific intervention on the subjects, these studies may be divided into two major groups: *observational* and *experimental* studies. Each of these designs has its own advantages and disadvantages. In observational studies, by definition, the researcher is looking for causes, predictors, and risk factors by observing the phenomenon without doing any intervention on the subjects, whereas in experimental designs, there is an intervention on the study subjects. In practice, however, there is no clear-cut distinction between different types of study design in biomedical research. *Case-only studies* are genetics studies in which individuals with and without a genotype of interest are compared, with an emphasis on environmental exposure.

For evaluation of etiologic and influencing role of some factors (such as risk factors) on the occurrence of diseases, we necessarily need to have a proper control group. Otherwise, no inferential decision can properly be made on the influencing role of risk factors. One of the study designs with control subjects is *case–control* study. These studies are designed to assess the cause–effect association by comparing a group of patients with a (matched) group of control subjects in terms of influencing or etiologic factors.

Some concerns in case–control studies, including control group and appropriate selection of control subjects, expensive cost for examining risk markers in both cases and controls (particularly in genetic studies), and the time-consuming process of such studies, have led to the development of the *case-only* method in studying the gene–environment interaction in human diseases. Investigators studying human malignancies have broadly used this method in recent years.

The case-only method was originally designed as a valid approach to the analysis and screening of genetic factors in the etiology of multifactorial diseases and also to assessing the gene–environment interactions in the etiology. In a case-only study, cases with and without the susceptible genotype are compared with each other in terms of the existence of the environmental exposure.

To conduct a case-only design, one applies the same epidemiological approaches to case selection rules as for any case–control study. The case-only study does not, however, have the complexity of rules for the selection of control subjects that usually appears in traditional case–control studies. The case-only method also requires fewer cases than the traditional case–control study. Furthermore, the case-only design has been reported to be

more efficient, precise, and powerful compared with a traditional case–control method.

Although the case-only design was originally created to improve the efficiency, power, and precision of the study of the gene–environment interactions by examining the prevalence of a specific genotype among case subjects only, it is now used to investigate how some other basic characteristics that vary slightly (or never vary) over time (e.g., gender, ethnicity, marital status, social and economic status) modify the effect of a time-dependent exposure (e.g., air pollution, extreme temperatures) on the outcome (e.g., myocardial infarction, death) in a group of cases only (e.g., decedents).

To avoid misinterpretation and bias, some technical assumptions should be taken into account in case-only studies. The assumption of independence between the susceptibility genotypes and the environmental exposures of interest in the population is the most important one that must be considered in conducting these studies. In practice, this assumption may be violated by some confounding factors (e.g., age, ethnic groups) if both exposure and genotype are affected. This assumption can be tested by some statistical methods. Some other technical considerations also must be assumed in the application of the case-only model in various studies of genetic factors. More details of these assumptions and the assessment of the gene–environment interaction in case-only studies can be found elsewhere.

*Saeed Dastgiri*

***See also*** Ethics in the Research Process

### Further Readings

Begg, C. B., & Zhang, Z. F. (1994). Statistical analysis of molecular epidemiology studies employing case series. *Cancer Epidemiology, Biomarkers & Prevention, 3,* 173–175.

Chatterjee, N., Kalaylioglu, Z., Shih, J. H., & Gail, M. H. (2006). Case-control and case-only designs with genotype and family history data: Estimating relative risk, residual familial aggregation, and cumulative risk. *Biometrics, 62,* 36–48.

Cheng, K. F. (2007). Analysis of case-only studies accounting for genotyping error. *Annals of Human Genetics, 71,* 238–248.

Greenland, S. (1993). Basic problems in interaction assessment. *Environmental Health Perspectives, 101*(Suppl. 4), 59–66.

Khoury, M. J., & Flanders, W. D. (1996). Nontraditional epidemiologic approaches in the analysis of gene-environment interaction: Case-control studies with no controls!; *American Journal of Epidemiology, 144,* 207–213.

Smeeth, L., Donnan, P. T., & Cook, D. G. (2006). The use of primary care databases: Case-control and case-only designs. *Family Practice, 23,* 597–604.

## Case Study

Case study research is a versatile approach to research in social and behavioral sciences. Case studies consist of detailed inquiry into a bounded entity or unit (or entities) in which the researcher either examines a relevant issue or reveals phenomena through the process of examining the entity within its social and cultural context. Case study has gained in popularity in recent years. However, it is difficult to define because researchers view it alternatively as a research design, an approach, a method, or even an outcome. This entry examines case study through different lenses to uncover the versatility in this type of research approach.

### Overview

Case study researchers have conducted studies in traditional disciplines such as anthropology, economics, history, political science, psychology, and sociology. Case studies have also emerged in areas such as medicine, law, nursing, business, administration, public policy, social work, and education. Case studies may be used as part of a larger study or as a stand-alone design. Case study research may be considered a method for inquiry or an evaluation of a bounded entity, program, or system. Case studies may consist of more than one entity (unit, thing) or of several cases within one entity, but care must be taken to limit the number of cases in order to allow for in-depth analysis and description of each case.

Researchers who have written about case study research have addressed it in different ways, depending on their perspectives and points of view.

Some researchers regard case study as a research process used to investigate a phenomenon in its real-world setting. Some have considered case study to be a design—a particular logic for setting up the study. Others think of case study as a qualitative approach to research that includes particular qualitative methods. Others have depicted it in terms of the final product, a written holistic examination, interpretation, and analysis of one or more entities or social units. Case study also has been defined in terms of the unit of study itself, or the entity being studied. Still other researchers consider that case study research encompasses all these notions taken together in relation to the research questions.

### *Case Study as a Bounded System*

Although scholars in various fields and disciplines have given case study research many forms, what these various definitions and perspectives have in common is the notion from Louis Smith that case study is inquiry about a *bounded* system. Case studies may be conducted about such entities as a single person or several persons, a single classroom or classrooms, a school, a program within a school, a business, an administrator, or a specific policy, and so on. The case study also may be about a complex, integrated system, as long as researchers are able to put boundaries or limits around the system being researched.

The notion of boundedness may be understood in more than one way. An entity is naturally bounded if the participants have come together by their own means for their own purposes having nothing to do with the research. An example would be a group of students in a particular classroom, a social club that meets on a regular basis, or the staff members in a department of a local business. The entity is naturally bounded because it consists of participants who are together for their own common purposes. A researcher is able to study the entity in its entirety for a time frame consistent with the research questions.

In other instances, an entity may be artificially bounded through the criteria set by a researcher. In this instance, the boundary is suggested by selecting from among participants to study an issue particular to some, but not all, the participants. For example, the researcher might study an issue that pertains only to a particular group of students, such as the study habits of sixth-grade students who are reading above grade level. In this situation, the case is artificially bounded because the researcher establishes the criteria for selection of the participants of the study as it relates to the issue being studied and the research questions being asked.

Selecting case study as a research design is appropriate for particular kinds of questions being asked. For example, if a researcher wants to know how a program works or why a program has been carried out in a particular way, case study can be of benefit. In other words, when a study is of an exploratory or explanatory nature, case study is well suited. In addition, case study works well for understanding processes because the researcher is able to get close to the participants within their local contexts. Case study design helps the researcher understand the complexity of a program or a policy, as well as its implementation and effects on the participants.

## Design Decisions

When a researcher begins case study research, a series of decisions must be made concerning the rationale, the design, the purpose, and the type of case study. In terms of the rationale for conducting the study, Robert Stake indicated three specific forms. Case study researchers may choose to learn about a case because of their inherent interest in the case itself. A teacher who is interested in following the work of a student who presents particular learning behaviors would conduct an *intrinsic* case study. In other words, the case is self-selected due to the inquirer's interest in the particular entity. Another example could be an evaluation researcher who may conduct an intrinsic case to evaluate a particular program. The program itself is of interest, and the evaluator is not attempting to compare that program to others.

On the other hand, an *instrumental* case is one that lends itself to the understanding of an issue or phenomenon beyond the case itself. For example, examining the change in teacher practices due to an educational reform may lead a researcher to select one teacher's classroom as a case, but with the intent of gaining a general understanding of the reform's effects on classrooms. In this instance, the

case selection is made for further understanding of a larger issue that may be instrumental in informing policy. For this purpose, one case may not be enough to fully understand the reform issue. The researcher may decide to have more than one case in order to see how the reform plays out in different classroom settings or in more than one school. If the intent is to study the reform in more than one setting, then the researcher would conduct a *collective* case study. Once the researcher has made a decision whether to understand the inherent nature of one case (intrinsic) or to understand a broader issue that may be represented by one of more cases (instrumental or collective), the researcher needs to decide the type of case that will illuminate the entity or issue in question.

### *Single-Case Design*

In terms of design, a researcher needs to decide whether to examine a single case, multiple cases, or several cases embedded within a larger system. In his work with case study research, Robert Yin suggests that one way to think about conducting a study of a single case is to consider it as a holistic examination of an entity that may demonstrate the tenets of a theory (i.e., *critical case*). If the case is highly unusual and warrants in-depth explanation, it would be considered an *extreme* or *unique* case. Another use for a single-case design would be a *typical* or *representative* case, in which the researcher is highlighting an everyday situation. The caution here is for the researcher to be well informed enough to know what is typical or commonplace. A *revelatory* case is one in which researchers are able to observe a phenomenon that had been previously inaccessible. One other type of single case is the *longitudinal* case, meaning one in which researchers can examine the same entity over time to see changes that occur.

Whether to conduct a holistic design or an embedded design depends on whether researchers are looking at issues related globally to one entity or whether subcases within that entity must be considered. For example, in studying the effects of an educational policy on a large, urban public school district, researchers could use either type of design, depending on what they deem important to examine. If researchers focused on the global aspect of the policy and how it may have changed the way the district structures its financial expenditures, the design would be holistic because the researcher would be exploring the entire school district to see effects. However, if the researcher examined the constraints placed on each regional subdivision within the district, each regional subdivision would become a subunit of the study as an embedded case design. Each regional office would be an embedded case and the district itself the major case. The researcher would procedurally examine each subunit and then reexamine the major case to see how the embedded subunits inform the whole case.

### *Multiple Case Design*

Multiple cases have been considered by some to be a separate type of study, but Yin considers them a variant of single-case design and thus similar in methodological procedures. This design also may be called a *collective case, cross-case,* or *comparative case* study. From the previous example, a team of researchers might consider studying the financial decisions made by the five largest school districts in the United States since the No Child Left Behind policy went into effect. Under this design, the research team would study each of the largest districts as an individual case, either alone or with embedded cases within each district.

After conducting the analysis for each district, the team would further conduct analyses across the five cases to see what elements they may have in common. In this way, researchers would potentially add to the understanding of the effects of policy implementation among these large school districts. By selecting only the five largest school districts, the researchers have limited the findings to cases that are relatively similar. If researchers wanted to compare the implementation under highly varied circumstances, they could use *maximum variation selection,* which entails finding cases with the greatest variation. For example, they may establish criteria for selecting a large urban school district, a small rural district, a medium-sized district in a small city, and perhaps a school district that serves a Native American population on a reservation. By deliberately selecting from among cases that had potential for differing implementation due to their distinct contexts,

the researchers expect to find as much variation as possible.

No matter what studies that involve more than a single case are called, what case studies have in common is that findings are presented as individual portraits that contribute to our understanding of the issues, first individually and then collectively. One question that often arises is whether the use of multiple cases can represent a form of generalizability, in that researchers may be able to show similarities of issues across the cases. The notion of generalizability in an approach that tends to be qualitative in nature may be of concern because of the way in which one selects the cases and collects and analyzes the data. For example, in statistical studies, the notion of generalizability comes from the form of sampling (e.g., randomized sampling to represent a population and a control group) and the types of measurement tools used (e.g., surveys that use Likert-type scales and therefore result in numerical data). In case study research, however, Yin has offered the notion that the different cases are similar to multiple experiments in which the researcher selects among similar and sometimes different situations to verify results. In this way, the cases become a form of generalizing to a theory, either taken from the literature or uncovered and grounded in the data.

### *Nature of the Case Study*

Another decision to be made in the design of a study is whether the purpose is primarily descriptive, exploratory, or explanatory. The nature of a descriptive case study is one in which the researcher uses *thick description* about the entity being studied so that the reader has a sense of having "been there, done that," in terms of the phenomenon studied within the context of the research setting. Exploratory case studies are those in which the research questions tend to be of the *what can be learned about this issue* type. The goal of this kind of study is to develop working hypotheses about the issue and perhaps to propose further research. An explanatory study is more suitable for delving into *how and why* things are happening as they are, especially if the events and people involved are to be observed over time.

## Case Study Approaches to Data

### *Data Collection*

One of the reasons that case study is such a versatile approach to research is that both quantitative and qualitative data may be used in the study, depending on the research questions asked. While many case studies have qualitative tendencies, due to the nature of exploring phenomena in context, some case study researchers use surveys to find out demographic and self-report information as part of the study. Researchers conducting case studies tend also to use interviews, direct observations, participant observation, and source documents as part of the data analysis and interpretation. The documents may consist of archival records, artifacts, and websites that provide information about the phenomenon in the context of what people make and use as resources in the setting.

### *Data Analysis*

The analytic role of the researcher is to systematically review these data, first making a detailed description of the case and the setting. It may be helpful for the researcher to outline a chronology of actions and events, although in further analysis, the chronology may not be as critical to the study as the thematic interpretations. However, it may be useful in terms of organizing what may otherwise be an unwieldy amount of data.

In further analysis, the researcher examines the data, one case at a time if multiple cases are involved, for patterns of actions and instances of issues. The researcher first notes what patterns are constructed from one set of data within the case and then examines subsequent data collected within the first case to see whether the patterns are consistent. At times, the patterns may be evident in the data alone, and at other times, the patterns may be related to relevant studies from the literature. If multiple cases are involved, a cross-case analysis is then conducted to find what patterns are consistent and under what conditions other patterns are apparent.

### *Reporting the Case*

While case study research has no predetermined reporting format, Stake has proposed an approach

for constructing an outline of the report. He suggests that the researcher start the report with a vignette from the case to draw the reader into time and place. In the next section, the researcher indentifies the issue studied and the methods for conducting the research. The next part is a full description of the case and context. Next, in describing the issues of the case, the researcher can build the complexity of the study for the reader. The researcher uses evidence from the case and may relate that evidence to other relevant research. In the next section, the researcher presents the *so what* of the case—a summary of claims made from the interpretation of data. At this point, the researcher may end with another vignette that reminds the reader of the complexity of the case in terms of realistic scenarios that readers may then use as a form of transference to their own settings and experiences.

### Versatility

Case study research demonstrates its utility in that the researcher can explore single or multiple phenomena or multiple examples of one phenomenon. The researcher can study one bounded entity in a holistic fashion or multiple subunits embedded within that entity. Through a case study approach, researchers may explore particular ways that participants conduct themselves in their localized contexts, or the researchers may choose to study processes involving program participants. Case study research can shed light on the particularity of a phenomenon or process while opening up avenues of understanding about the entity or entities involved in those processes.

*LeAnn Grogan Putney*

#### Further Readings

Creswell, J. W. (2007). *Qualitative inquiry and research design: Choosing among five approaches*. Thousand Oaks, CA: Sage.

Creswell, J. W., & Maietta, R. C. (2002). Qualitative research. In D. C. Miller & N. J. Salkind (Eds.), *Handbook of research design and social measurement* (pp. 162–163). Thousand Oaks, CA: Sage.

Gomm, R., Hammersley, M., & Foster, P. (2000). *Case study method: Key issues, Key texts*. Thousand Oaks, CA: Sage.

Merriam, S. B. (1998). *Qualitative research and case study applications in education*. San Francisco: Jossey-Bass.

Stake, R. E. (1995). *Art of case study research*. Thousand Oaks, CA: Sage.

Stake, R. E. (2005). *Multiple case study analysis*. New York: Guilford.

Yin, R. K. (2003). *Case study research*. Thousand Oaks, CA: Sage.

## Categorical Data Analysis

A categorical variable consists of a set of nonoverlapping categories. Categorical data are *counts* for those categories. The measurement scale of a categorical variable is *ordinal* if the categories exhibit a natural ordering, such as opinion variables with categories from "strongly disagree" to "strongly agree." The measurement scale is *nominal* if there is no inherent ordering. The types of possible analysis for categorical data depend on the measurement scale.

### Types of Analysis

When the subjects measured are cross-classified on two or more categorical variables, the table of counts for the various combinations of categories is a *contingency table*. The information in a contingency table can be summarized and further analyzed through appropriate *measures of association* and models as discussed below. These measures and models differentiate according to the nature of the classification variables (nominal or ordinal).

Most studies distinguish between one or more *response variables* and a set of *explanatory variables*. When the main focus is on the association and interaction structure among a set of response variables, such as whether two variables are conditionally independent given values for the other variables, *loglinear models* are useful, as described in a later section. More commonly, research questions focus on effects of explanatory variables on a categorical response variable. Those explanatory variables might be categorical, quantitative, or of both types. *Logistic regression models* are then of

particular interest. Initially such models were developed for *binary* (success–failure) response variables. They describe the *logit,* which is $\log[P(Y = 1)/P(Y = 2)]$, using the equation

$$\log[P(Y = 1)/P(Y = 2)] = a + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p,$$

where $Y$ is the binary response variable and $x_1, \ldots, x_p$ the set of the explanatory variables. The logistic regression model was later extended to nominal and ordinal response variables. For a nominal response $Y$ with $J$ categories, the model simultaneously describes

$$\log[P(Y = 1)/P(Y = J)],$$
$$\log[P(Y = 2)/P(Y = J)], \ldots,$$
$$\log[P(Y = J - 1)/P(Y = J)].$$

For ordinal responses, a popular model uses explanatory variables to predict a logit defined in terms of a cumulative probability,

$$\log[P(Y \le j)/P(Y > j)], j = 1, 2, \ldots, J - 1.$$

For categorical data, the binomial and multinomial distributions play the central role that the normal does for quantitative data. Models for categorical data assuming the binomial or multinomial were unified with standard regression and *analysis of variance* (ANOVA) models for quantitative data assuming normality were unified through the introduction of the *generalized linear model* (GLM). This very wide class of models can incorporate data assumed to come from any of a variety of standard distributions (such as the normal, binomial, and Poisson). The GLM relates a function of the mean (such as the log or logit of the mean) to explanatory variables with a linear predictor. Certain GLMs for counts, such as *Poisson regression* models, relate naturally to log linear and logistic models for binomial and multinomial responses.

More recently, methods for categorical data have been extended to include *clustered data,* for which observations within each cluster are allowed to be correlated. A very important special case is that of *repeated measurements,* such as in a longitudinal study in which each subject provides a cluster of observations taken at different times. One way this is done is to introduce a *random effect* in the model to represent each cluster, thus extending the GLM to a *generalized linear mixed model,* the *mixed* referring to the model's containing both random effects and the usual sorts of fixed effects.

## Two-Way Contingency Tables

Two categorical variables are *independent* if the probability of response in any particular category of one variable is the same for each category of the other variable. The most well-known result on two-way contingency tables is the test of the null hypothesis of independence, introduced by Karl Pearson in 1900. If $X$ and $Y$ are two categorical variables with $I$ and $J$ categories, respectively, then their cross-classification leads to a $I \times J$ table of observed frequencies $\mathbf{n} = (n_{ij})$. Under this hypothesis, the expected cell frequencies are values that have the same marginal totals as the observed counts but perfectly satisfy the hypothesis. They equal $m_{ij} = n\pi_{i+}\pi_{+j}$, $i = 1, \ldots, I$, $j = 1, \ldots, J$, where $n$ is the total sample size ($n = \sum_{i,j} n_{ij}$) and $\pi_{i+}(\pi_{+j})$ is the $i$th row ($j$th column) marginal of the underlying probabilities matrix $\boldsymbol{\pi} = (\pi_{ij})$. Then the corresponding maximum likelihood (ML) estimates equal $\hat{m}_{ij} = np_{i+}p_{+j} = \frac{n_{i+}n_{+j}}{n}$, where $p_{ij}$ denotes the sample proportion in cell $(i, j)$. The hypothesis of independence is tested through Pearson's chi-square statistic,

$$X^2 = \sum_{i,j} \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}}. \qquad (1)$$

The $p$ value is the right-tail probability above the observed $X^2$ value. The distribution of $X^2$ under the null hypothesis is approximated by a $\chi^2_{(I-1)(J-1)}$, provided that the individual expected cell frequencies are not too small. In fact, Pearson claimed that the associated degrees of freedom ($df$) were $IJ - 1$, and R. A. Fisher corrected this in 1922. Fisher later proposed a small-sample test of independence for $2 \times 2$ tables, now referred to as *Fisher's exact test.* This test was later extended to $I \times J$ tables as well as to more complex hypotheses in both two-way and multiway tables. When a contingency table has ordered row or column categories (ordinal variables), specialized methods can take advantage of that ordering.

Ultimately more important than mere testing of significance is the estimation of the strength of the association. For ordinal data, measures can incorporate information about the direction (positive or negative) of the association as well.

More generally, models can be formulated that are more complex than independence, and expected frequencies $m_{ij}$ can be estimated under the constraint that the model holds. If $\hat{m}_{ij}$ are the corresponding maximum likelihood estimates, then, to test the hypothesis that the model holds, one can use the Pearson statistic (Equation 1) or the statistic that results from the standard statistical approach of conducting a *likelihood-ratio test*, which is

$$G^2 = 2\sum_{i,j} n_{ij} \ln\left(\frac{n_{ij}}{\hat{m}_{ij}}\right). \qquad (2)$$

Under the null hypothesis, both statistics have the same large-sample chi-square distribution.

The special case of the $2 \times 2$ table occurs commonly in practice, for instance for comparing two groups on a success/fail–type outcome. In a $2 \times 2$ table, the basic measure of association is the *odds ratio*. For the probability table

$$\begin{bmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \end{bmatrix}$$

the odds ratio is defined as $\theta = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}$. Independence corresponds to $\theta = 1$. Inference about the odds ratio can be based on the fact that for large samples,

$$\log(\hat{\theta}) \sim N\left(\log(\theta), \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}\right).$$

The odds ratio relates to the *relative risk r*. In particular, if we assume that the rows of the above $2 \times 2$ table represent two independent groups of subjects (A and B) and the columns correspond to presence/absence of a disease, then the relative risk for this disease is defined as $r = \frac{\pi_A}{\pi_B}$, where $\pi_A = \frac{\pi_{11}}{\pi_{1+}}$ is the probability of disease for the first group and $\pi_B$ is defined analogously. Since $\theta = r\frac{1-\pi_B}{1-\pi_A}$, it follows that $\theta \approx r$ whenever $\pi_A$ and $\pi_B$ are close to 0.

## Models for Two-Way Contingency Tables

Independence between the classification variables $X$ and $Y$ (i.e., $m_{ij} = n\pi_{i+}\pi_{+j}$, for all $i$ and $j$) can equivalently be expressed in terms of a log linear model as

$$\log(m_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y,$$
$$i = 1, \ldots, I,\ j = 1, \ldots, J.$$

The more general model that allows association between the variables is

$$\log(m_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY},$$
$$i = 1, \ldots, I,\ \ j = 1, \ldots, J. \qquad (3)$$

Loglinear models describe the way the categorical variables and their association influence the count in each cell of the contingency table. They can be considered as a discrete analogue of ANOVA. The two-factor interaction terms relate to odds ratios describing the association. As in ANOVA models, some parameters are redundant in these specifications, and software reports estimates by assuming certain constraints.

The general model (Equation 3) does not impose any structure on the underlying association, and so it fits the data perfectly. Associations can be modeled through *association models*. The simplest such model, the *linear-by-linear association model*, is relevant when both classification variables are ordinal. It replaces the interaction term $\lambda_{ij}^{XY}$ by the product $\phi\mu_i\nu_j$, where $\mu_i$ and $\nu_j$ are known scores assigned to the row and column categories, respectively. This model,

$$\log(m_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \phi\mu_i\nu_j,$$
$$i = 1, \ldots, I,\ \ j = 1, \ldots, J, \qquad (4)$$

has only one parameter more than the independence model, namely $\phi$. Consequently, the associated *df* are $(I-1)(J-1)-1$, and once it holds, independence can be tested conditionally on it by testing $\phi = 0$ via a more powerful test with $df = 1$. The linear-by-linear association model (Equation 4) can equivalently be expressed in terms of the $(I-1)(J-1)$ local odds ratios $\theta_{ij} = \frac{\pi_{ij}\pi_{i+1,j+1}}{\pi_{i,j+1}\pi_{i+1,j}} (I = 1, \ldots, I-1, j = 1, \ldots, J-1)$, defined by adjacent rows and columns of the table:

$$\theta_{ij} = \exp[\phi(\mu_{i+1} - \mu_i)(\nu_{j+1} - \nu_j)], \quad i = 1, \ldots, I-1, \; j = 1, \ldots, J-1. \tag{5}$$

With equally spaced scores, all the local odds ratios are identical, and the model is referred to as *uniform association*. More general models treat one or both sets of scores as parameters. Association models have been mainly developed by L. Goodman.

Another popular method for studying the pattern of association between the row and column categories of a two-way contingency table is *correspondence analysis* (CA). It is mainly a descriptive method. CA assigns optimal scores to the row and column categories and plots these scores in two or three dimensions, providing thus a reduced rank display of the underlying association.

The special case of square $I \times I$ contingency tables with the same categories for the rows and the columns occurs with matched-pairs data. For example, such tables occur in the study of *rater agreement* and in the analysis of social mobility. A condition of particular interest for such data is *marginal homogeneity*, that $\pi_{i+} = \pi_{+i}$, $i = 1, \ldots, I$. For the $2 \times 2$ case of binary matched pairs, the test comparing the margins using the chi-square statistic $(n_{12} - n_{21})^2/(n_{12} + n_{21})$ is called *McNemar's test*.

## Multiway Contingency Tables

The models described earlier for two-way tables extend to higher dimensions. For multidimensional tables, a number of models are available, varying in terms of the complexity of the association structure. For three variables, for instance, models include ones for which (a) the variables are mutually independent; (b) two of the variables are associated but are jointly independent of the third; (c) two of the variables are conditionally independent, given the third variable, but may both be associated with the third; and (d) each pair of variables is associated, but the association between each pair has the same strength at each level of the third variable. Because the number of possible models increases dramatically with the dimension, model selection methods become more important as the dimension of the table increases. When the underlying theory for the research study does not suggest particular methods, one can use the same methods that are available for ordinary regression models, such as stepwise selection methods and fit indices such as Akaike Information Criterion.

Loglinear models for multiway tables can include higher order interactions up to the order equal to the dimension of the table. Two-factor terms describe conditional association between two variables, three-factor terms describe how the conditional association varies among categories of a third variable, and so forth. CA has also been extended to higher dimensional tables, leading to *multiple CA*.

Historically, a common way to analyze higher way contingency tables was to analyze all the two-way tables obtained by collapsing the table over the other variables. However, the two-way associations can be quite different from conditional associations in which other variables are controlled. The association can even change direction, a phenomenon known as *Simpson's paradox*. Conditions under which tables can be collapsed are most easily expressed and visualized using graphical models that portray each variable as a node and a conditional association as a connection between two nodes. The patterns of associations and their strengths in two-way or multiway tables can also be illustrated through special plots called *mosaic plots*.

## Inference and Software

Least squares is not an optimal estimation method for categorical data, because the variance of sample proportions is not constant but rather depends on the corresponding population proportions. Because of this, parameters for categorical data were estimated historically by the use of *weighted least squares*, giving more weight to observations having smaller variances. Currently, the most popular estimation method is *maximum likelihood*, which is an optimal method for large samples for any type of data. The *Bayesian* approach to inference, in which researchers combine the information from the data with their prior beliefs to obtain posterior distributions for the parameters of interest, is becoming more popular. For large samples, all these methods yield similar results.

Standard statistical packages, such as SAS, SPSS (an IBM company, formerly called PASW® Statistics), Stata, and S-Plus and R, are well suited for analyzing categorical data. Such packages now

have facility for fitting GLMs, and most of the standard methods for categorical data can be viewed as special cases of such modeling. Bayesian analysis of categorical data can be carried out through WINBUGS. Specialized software such as the programs StatXact and LogXact, developed by Cytel Software, are available for small-sample exact methods of inference for contingency tables and for logistic regression parameters.

*Maria Kateri and Alan Agresti*

***See also*** Categorical Variable; Correspondence Analysis; General Linear Model; R; SAS; Simpson's Paradox; SPSS

**Further Readings**

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York: Wiley.
Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed.). New York: Wiley.
Agresti, A. (2010). *Analysis of Ordinal Categorical Data*. New York: Wiley.
Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge: MIT Press.
Congdon, P. (2005). *Bayesian Models for Categorical Data*. New York: Wiley.
Goodman, L. A. (1986). Some useful extensions of the usual correspondence analysis and the usual log-linear models approach in the analysis of contingency tables with or without missing entries. *International Statistical Review, 54*, 243-309.
Kateri, M. (2008). Categorical data. In S. Kotz (Ed.), *Encyclopedia of statistical sciences* (2nd ed.). Hoboken, NJ: Wiley-Interscience.

**Websites**

Cytel Software: http://www.cytel.com
WINBUGS: http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml

## Categorical Variable

Categorical variables are qualitative data in which the values are assigned to a set of distinct groups or categories. These groups may consist of alphabetic (e.g., male, female) or numeric labels (e.g., male = 0, female = 1) that do not contain mathematical information beyond the frequency counts related to group membership. Instead, categorical variables often provide valuable social-oriented information that is not quantitative by nature (e.g., hair color, religion, ethnic group).

In the hierarchy of measurement levels, categorical variables are associated with the two lowest variable classification orders, nominal or ordinal scales, depending on whether the variable groups exhibit an intrinsic ranking. A nominal measurement level consists purely of categorical variables that have no ordered structure for intergroup comparison. If the categories can be ranked according to a collectively accepted protocol (e.g., from lowest to highest), then these variables are ordered categorical, a subset of the ordinal level of measurement.

Categorical variables at the nominal level of measurement have two properties. First, the categories are mutually exclusive. That is, an object can belong to only one category. Second, the data categories have no logical order. For example, researchers can measure research participants' religious backgrounds, such as Jewish, Protestant, Muslim, and so on, but they cannot order these variables from lowest to highest. It should be noted that when categories get numeric labels such as male = 0 and female = 1 or control group = 0 and treatment group = 1, the numbers are merely labels and do not indicate one category is "better" on some aspect than another. The numbers are used as symbols (codes) and do not reflect either quantities or a rank ordering. *Dummy coding* is the quantification of a variable with two categories (e.g., boys, girls). Dummy coding will allow the researcher to conduct specific analyses such as the *point-biserial correlation coefficient,* in which a dichotomous categorical variable is related to a variable that is continuous. One example of the use of point-biserial correlation is to compare males with females on a measure of mathematical ability.

Categorical variables at the ordinal level of measurement have the following properties: (a) the data categories are mutually exclusive, (b) the data categories have some logical order, and (c) the data categories are scaled according to the amount of a particular characteristic. Grades in courses (i.e., A, B, C, D, and F) are an example. The person who earns an A in a course has a higher level of achievement than one who gets a B, according to

the criteria used for measurement by the course instructor. However, one cannot assume that the difference between an A and a B is the same as the difference between a B and a C. Similarly, researchers might set up a Likert-type scale to measure level of satisfaction with one's job and assign a 5 to indicate *extremely satisfied,* 4 to indicate *very satisfied,* 3 to indicate *moderately satisfied,* and so on. A person who gives a rating of 5 feels more job satisfaction than a person who gives a rating of 3, but it has no meaning to say that one person has 2 units more satisfaction with a job than another has or exactly how much more satisfied one is with a job than another person is.

In addition to verbal descriptions, categorical variables are often presented visually using tables and charts that indicate the group frequency (i.e., the number of values in a given category). Contingency tables show the number of counts in each category and increase in complexity as more attributes are examined for the same object. For example, a car can be classified according to color, manufacturer, and model. This information can be displayed in a contingency table showing the number of cars that meet each of these characteristics (e.g., the number of cars that are white and manufactured by General Motors). This same information can be expressed graphically using a bar chart or pie chart. Bar charts display the data as elongated bars with lengths proportional to category frequency, with the category labels typically being the x-axis and the number of values the y-axis. On the other hand, pie charts show categorical data as proportions of the total value or as a percentage or fraction. Each category constitutes a section of a circular graph or "pie" and represents a subset of the 100% or fractional total. In the car example, if 25 cars out of a sample of 100 cars were white, then 25%, or one quarter, of the circular pie chart would be shaded, and the remaining portion of the chart would be shaded alternative colors based on the remaining categorical data (i.e., cars in colors other than white).

Specific statistical tests that differ from other quantitative approaches are designed to account for data at the categorical level. The only *measure of central tendency* appropriate for categorical variables at the nominal level is *mode* (the most frequent category or categories if there is more than one mode), but at the ordinal level, the median or point below which 50% of the scores fall is also used. The chi-square distribution is used for categorical data at the nominal level. Observed frequencies in each category are compared with the theoretical or expected frequencies. Types of correlation coefficients that use categorical data include point biserial; Spearman rho, in which both variables are at the ordinal level; and phi, in which both variables are dichotomous (e.g., boys vs. girls on a yes–no question). Categorical variables can also be used in various statistical analyses such as *t* tests, analysis of variance, multivariate analysis of variance, simple and multiple regression analysis, and discriminant analysis.

*Karen D. Multon and Jill S. M. Coleman*

***See also*** Bar Chart; Categorical Data Analysis; Chi-Square Test; Levels of Measurement; Likert Scaling; Nominal Scale; Ordinal Scale; Pie Chart; Variable

**Further Readings**

Siegel, A. F., & Morgan, C. J. (1996). *Statistics and data analysis: An introduction* (2nd ed.). New York: Wiley.
Simonoff, J. S. (2003). *Analyzing categorical data.* New York: Springer.

## Causal-Comparative Design

A causal-comparative design is a research design that seeks to find relationships between independent and dependent variables after an action or event has already occurred. The researcher's goal is to determine whether the independent variable affected the outcome, or dependent variable, by comparing two or more groups of individuals. There are similarities and differences between causal-comparative research, also referred to as *ex post facto research,* and both correlational and experimental research. This entry discusses these differences, as well as the benefits, process, limitations, and criticism of this type of research design. To demonstrate how to use causal-comparative research, examples in education are presented.

## Comparisons With Correlational Research

Many similarities exist between causal-comparative research and correlational research. Both methods are useful when experimental research has been deemed impossible or unethical as the research design for a particular question. Both causal-comparative and correlational research designs attempt to determine relationships among variables, but neither allows for the actual manipulation of these variables. Thus, neither can definitively state that a true cause-and-effect relationship occurred between these variables. Finally, neither type of design randomly places subjects into control and experimental groups, which limits the generalizability of the results.

Despite similarities, there are distinct differences between causal-comparative and correlational research designs. In causal-comparative research, the researcher investigates the effect of an independent variable on a dependent variable by comparing two or more groups of individuals. For example, an educational researcher may want to determine whether a computer-based ACT program has a positive effect on ACT test scores. In this example, the researcher would compare the ACT scores from a group of students that completed the program with scores from a group that did not complete the program. In correlational research, the researcher works with only one group of individuals. Instead of comparing two groups, the correlational researcher examines the effect of one or more independent variables on the dependent variable within the same group of subjects. Using the same example as above, the correlational researcher would select one group of subjects who have completed the computer-based ACT program. The researcher would use statistical measures to determine whether there was a positive relationship between completion of the ACT program and the students' ACT scores.

## Comparisons With Experimental Research

A few aspects of causal-comparative research parallel experimental research designs. Unlike correlational research, both experimental research and causal-comparative research typically compare two or more groups of subjects. Research subjects are generally split into groups on the basis of the independent variable that is the focus of the study. Another similarity is that the goal of both types of research is to determine what effect the independent variable may or may not have on the dependent variable or variables.

While the premises of the two research designs are comparable, there are vast differences between causal-comparative research and experimental research. First and foremost, causal-comparative research occurs after the event or action has been completed. It is a retrospective way of determining what may have caused something to occur. In true experimental research designs, the researcher manipulates the independent variable in the experimental group. Because the researcher has more control over the variables in an experimental research study, the argument that the independent variable caused the change in the dependent variable is much stronger. Another major distinction between the two types of research is random sampling. In causal-comparative research, the research subjects are already in groups because the action or event has already occurred, whereas subjects in experimental research designs are randomly selected prior to the manipulation of the variables. This allows for wider generalizations to be made from the results of the study.

Table 1 breaks down the causal-comparative, correlational, and experimental methods in reference to whether each investigates cause-effect and whether the variables can be manipulated. In addition, it notes whether groups are randomly assigned and whether the methods study groups or individuals.

## When to Use Causal-Comparative Research Designs

Although experimental research results in more compelling arguments for causation, there are many times when such research cannot, or should not, be conducted. Causal-comparative research provides a viable form of research that can be conducted when other methods will not work. There are particular independent variables that are not capable of being manipulated, including gender, ethnicity, socioeconomic level, education level, and religious preferences. For instance, if researchers intend to examine whether ethnicity affects

**Table 1** Comparison of Causal-Comparative, Correlational, and Experimental Research

| *Method* | *Investigates Cause–Effect* | *Manipulates Variable* | *Randomly Assigns Participants to Groups* | *Involves Group Comparisons* | *Studies Groups or Individuals* | *Focus* | *Identifies Variables for Experimental Exploration* |
|---|---|---|---|---|---|---|---|
| Causal-comparative research | Yes | No (it already occurred) | No (groups formed prior to study) | Yes | Two or more groups of individuals and one independent variable | Focus on differences of variables between groups | Yes |
| Correlational research | No | No | No (only one group) | No | Two or more variables and one group of individuals | Focus on relationship among variables | Yes |
| Experimental research | Yes | Yes | Yes | Yes | Groups or individuals depending on design | Depends on design; focuses on cause/effect of variables | Yes |

self-esteem in a rural high school, they cannot manipulate a subject's ethnicity. This independent variable has already been decided, so the researchers must look to another method of determining cause. In this case, the researchers would group students according to their ethnicity and then administer self-esteem assessments. Although the researchers may find that one ethnic group has higher scores than another, they must proceed with caution when interpreting the results. In this example, it might be possible that one ethnic group is also from a higher socioeconomic demographic, which may mean that the socioeconomic variable affected the assessment scores.

Some independent variables should not be manipulated. In educational research, for example, ethical considerations require that the research method not deny potentially useful services to students. For instance, if a guidance counselor wanted to determine whether advanced placement course selection affected college choice, the counselor could not ethically force some students to take certain classes and prevent others from taking the same classes. In this case, the counselor could still compare students who had completed advanced placement courses with those who had not, but causal conclusions are more difficult than with an experimental design.

Furthermore, causal-comparative research may prove to be the design of choice even when experimental research is possible. Experimental research is both time-consuming and costly. Many school districts do not have the resources to conduct a full-scale experimental research study, so educational leaders may choose to do a causal-comparative study. For example, the leadership might want to determine whether a particular math curriculum would improve math ACT scores more effectively than the curriculum already in place in the school district. Before implementing the new curriculum throughout the district, the school leaders might conduct a causal-comparative study, comparing their district's math ACT scores with those from a school district that has already used the curriculum. In addition, causal-comparative research is often selected as a precursor to experimental research. In the math curriculum example, if the causal-comparative study demonstrates that the curriculum has a positive effect on student math ACT scores, the school leaders may then choose to conduct a full experimental research study by piloting the curriculum in one of the schools in the district.

## Conducting Causal Comparative Research

The basic outline for conducting causal comparative research is similar to that of other research

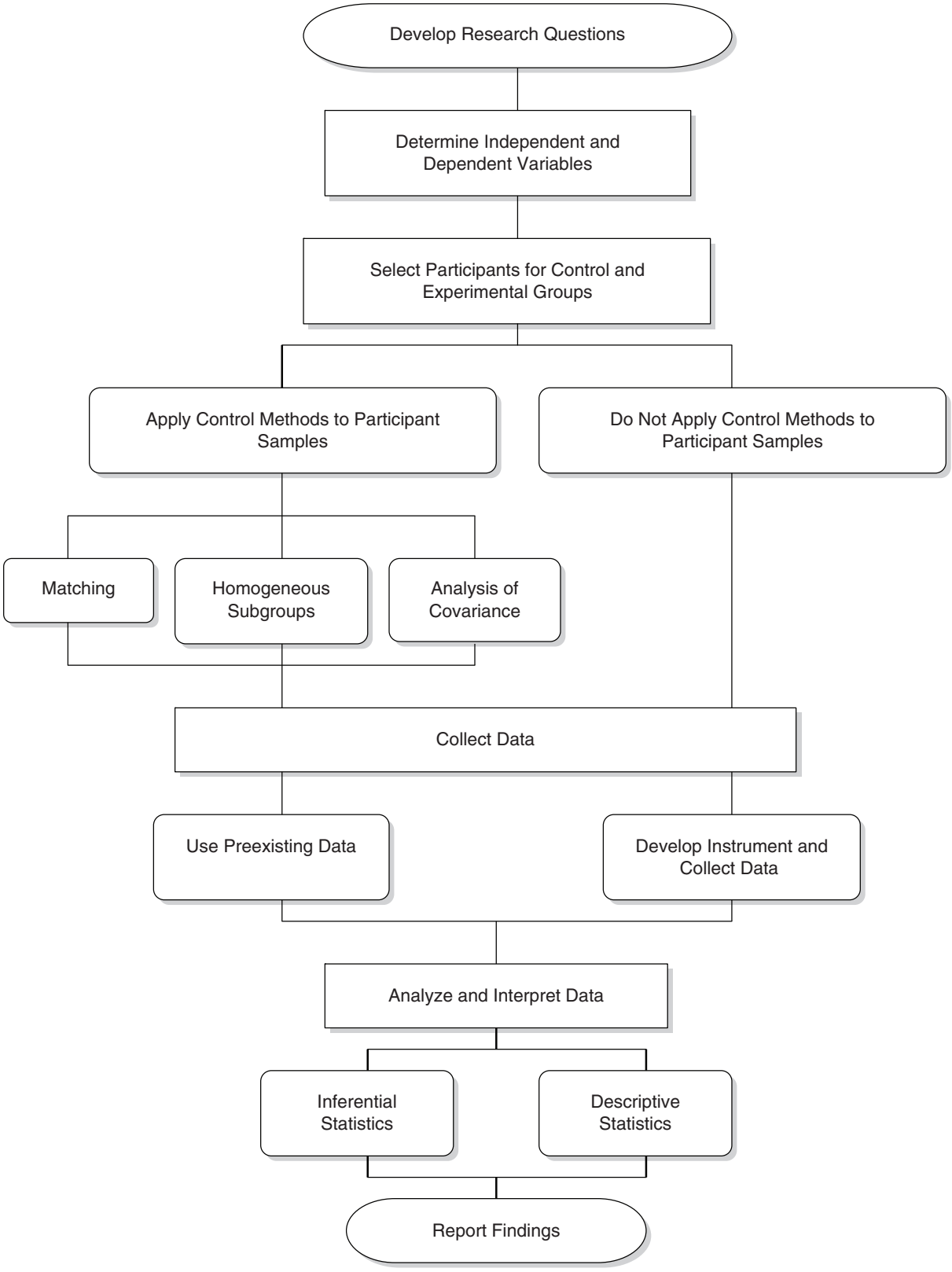



**Figure 1** Flowchart for Conducting a Study

designs. Once the researcher determines the focus of the research and develops hypotheses, he or she selects a sample of participants for both an experimental and a control group. Depending on the type of sample and the research question, the researcher may measure potentially confounding variables to include them in eventual analyses. The next step is to collect data. The researcher then analyzes the data, interprets the results, and reports the findings. Figure 1 illustrates this process.

### *Determine the Focus of Research*

As in other research designs, the first step in conducting a causal-comparative research study is to identify a specific research question and generate a hypothesis. In doing so, the researcher identifies a dependent variable, such as high dropout rates in high schools. The next step is to explore reasons the dependent variable has occurred or is occurring. In this example, several issues may affect dropout rates, including such elements as parental support, socioeconomic level, gender, ethnicity, and teacher support. The researcher will need to select which issue is of importance to his or her research goals. One hypothesis might be, "Students from lower socioeconomic levels drop out of high school at higher rates than students from higher socioeconomic levels." Thus, the independent variable in this scenario would be socioeconomic levels of high school students.

It is important to remember that many factors affect dropout rates. Controlling for such factors in causal-comparative research is discussed later in this entry. Once the researcher has identified the main research problem, he or she operationally defines the variables. In the above hypothesis, the dependent variable of high school dropout rates is fairly self-explanatory. However, the researcher would need to establish what constitutes lower socioeconomic levels and higher socioeconomic levels. The researcher may also wish to clarify the target population, such as what specific type of high school will be the focus of the study. Using the above example, the final research question might be, "Does socioeconomic status affect dropout rates in the Appalachian rural high schools in East Tennessee?" In this case, causal comparative would be the most appropriate method of research because the independent variable of socioeconomic status cannot be manipulated.

Because many factors may influence the dependent variable, the researcher should be aware of, and possibly test for, a variety of independent variables. For instance, if the researcher wishes to determine whether socioeconomic level affects a student's decision to drop out of high school, the researcher may also want to test for other potential causes, such as parental support, academic ability, disciplinary issues, and other viable options. If other variables can be ruled out, the case for socioeconomic level's influencing the dropout rate will be much stronger.

### *Participant Sampling and Threats to Internal Validity*

In causal-comparative research, two or more groups of participants are compared. These groups are defined by the different levels of the independent variable(s). In the previous example, the researcher compares a group of high school dropouts with a group of high school students who have not dropped out of school. Although this is not an experimental design, causal-comparative researchers may still randomly select participants within each group. For example, a researcher may select every fifth dropout and every fifth high school student. However, because the participants are not randomly selected and placed into groups, internal validity is threatened. To strengthen the research design and counter threats to internal validity, the researcher might choose to impose the selection techniques of matching, using homogeneous subgroups, or analysis of covariance (ANCOVA), or both.

#### Matching

One method of strengthening the research sample is to select participants by matching. Using this technique, the researcher identifies one or more characteristics and selects participants who have these characteristics for both the control and the experimental groups. For example, if the researcher wishes to control for gender and grade level, he or she would ensure that both groups matched on these characteristics. If a male 12th-grade student is selected for the experimental

group, then a male 12th-grade student must be selected for the control group. In this way the researcher is able to control these two extraneous variables.

#### Comparing Homogeneous Subgroups

Another control technique used in causal-comparative research is to compare subgroups that are clustered according to a particular variable. For example, the researcher may choose to group and compare students by grade level. He or she would then categorize the sample into subgroups, comparing 9th-grade students with other 9th-grade students, 10th-grade students with other 10th-grade students, and so forth. Thus, the researcher has controlled the sample for grade level.

#### Analysis of Covariance

Using the ANCOVA statistical method, the researcher is able to adjust previously disproportionate scores on a pretest in order to equalize the groups on some covariate (control variable). The researcher may want to control for ACT scores and their impact on high school dropout rates. In comparing the groups, if one group's ACT scores are much higher or lower than the other's, the researcher may use the technique of ANCOVA to balance the two groups. This technique is particularly useful when the research design includes a pretest, which assesses the dependent variable before any manipulation or treatment has occurred. For example, to determine the effect of an ACT curriculum on students, a researcher would determine the students' baseline ACT scores. If the control group had scores that were much higher to begin with than the experimental group's scores, the researcher might use the ANCOVA technique to balance the two groups.

### *Instrumentation and Data Collection*

The methods of collecting data for a causal-comparative research study do not differ from any other method of research. Questionnaires, pretests and posttests, various assessments, and behavior observation are common methods for collecting data in any research study. It is important, however, to also gather as much demographic information as possible, especially if the researcher is planning to use the control method of matching.

### *Data Analysis and Interpretation*

Once the data have been collected, the researcher analyzes and interprets the results. Although causal-comparative research is not true experimental research, there are many methods of analyzing the resulting data, depending on the research design. It is important to remember that no matter what methods are used, causal-comparative research does not definitively prove cause-and-effect results. Nevertheless, the results will provide insights into causal relationships between the variables.

#### Inferential Statistics

When using inferential statistics in causal-comparative research, the researcher hopes to demonstrate that a relationship exists between the independent and dependent variables. Again, the appropriate method of analyzing data using this type of statistics is determined by the design of the research study. The three most commonly used methods for causal-comparative research are the chi-square test, paired-samples and independent *t* tests, and analysis of variance (ANOVA) or ANCOVA.

Pearson's chi-square, the most commonly used chi-square test, allows the researcher to determine whether there is a statistically significant relationship between the experimental and control groups based on frequency counts. This test is useful when the researcher is working with nominal data, that is, different categories of treatment or participant characteristics, such as gender. For example, if a researcher wants to determine whether males and females learn more efficiently from different teaching styles, the researcher may compare a group of male students with a group of female students. Both groups may be asked whether they learn better from audiovisual aids, group discussion, or lecture. The researcher could use chi-square testing to analyze the data for evidence of a relationship.

Another method of testing relationships in causal-comparative research is to use independent or dependent *t* tests. When the researcher is

comparing the mean scores of two groups, these tests can determine whether there is a significant difference between the control and experimental groups. The independent *t* test is used in research designs when no controls have been applied to the samples, while the dependent *t* test is appropriate for designs in which matching has been applied to the samples. One example of the use of *t* testing in causal-comparative research is to determine the significant difference in math course grades between two groups of elementary school students when one group has completed a math tutoring course. If the two samples were matched on certain variables such as gender and parental support, the dependent *t* test would be used. If no matching was involved, the independent *t* test would be the test of choice. The results of the *t* test allow the researcher to determine whether there is a statistically significant relationship between the independent variable of the math tutoring course and the dependent variable of math course grade.

To test for relationships between three or more groups and a continuous dependent variable, a researcher might select the statistical technique of one-way ANOVA. Like the independent *t* test, this test determines whether there is a significant difference between groups based on their mean scores. In the example of the math tutoring course, the researcher may want to determine the effects of the course for students who attended daily sessions and students who attended weekly sessions, while also assessing students who never attended sessions. The researcher could compare the average math grades of the three groups to determine whether the tutoring course had a significant impact on the students' overall math grades.

## Limitations

Although causal-comparative research is effective in establishing relationships between variables, there are many limitations to this type of research. Because causal-comparative research occurs ex post facto, the researcher has no control over the variables and thus cannot manipulate them. In addition, there are often variables other than the independent variable(s) that may impact the dependent variable(s). Thus, the researcher cannot be certain that the independent variable caused the changes in the dependent variable. In order to counter this issue, the researcher must test several different theories to establish whether other variables affect the dependent variable. The researcher can reinforce the research hypothesis if he or she can demonstrate that other variables do not have a significant impact on the dependent variable.

Reversal causation is another issue that may arise in causal-comparative research. This problem occurs when it is not clear that the independent variable caused the changes in the dependent variable, or that a dependent variable caused the independent variable to occur. For example, if a researcher hoped to determine the success rate of an advanced English program on students' grades, he or she would have to determine whether the English program had a positive effect on the students, or in the case of reversal causation, whether students who make higher grades do better in the English program. In this scenario, the researcher could establish which event occurred first. If the students had lower grades before taking the course, then the argument that the course impacted the grades would be stronger.

The inability to construct random samples is another limitation in causal-comparative research. There is no opportunity to randomly choose participants for the experimental and control groups because the events or actions have already occurred. Without random assignment, the results cannot be generalized to the public, and thus the researcher's results are limited to the population that has been included in the research study. Despite this problem, researchers may strengthen their argument by randomly selecting participants from the previously established groups. For example, if there were 100 students who had completed a computer-based learning course, the researcher would randomly choose 20 students to compare with 20 randomly chosen students who had not completed the course. Another method of reinforcing the study would be to test the hypothesis with several different population samples. If the results are the same in all or most of the sample, the argument will be more convincing.

## Criticisms

There have been many criticisms of causal-comparative research. For the most part, critics reject the idea that causal-comparative research

results should be interpreted as evidence of causal relationships. These critics believe that there are too many limitations in this type of research to allow for a suggestion of cause and effect. Some critics are frustrated with researchers who hold that causal-comparative research provides stronger causal evidence than correlational research does. Instead, they maintain that neither type of research can produce evidence of a causal relationship, so neither is better than the other. Most of these critics argue that experimental research designs are the only method of research that can illustrate any type of causal relationships between variables. Almost all agree, however, that experimental designs potentially provide the strongest evidence for causation.

*Ernest W. Brewer and Jennifer Kuhn*

***See also*** Cause and Effect; Correlation; Experimental Design; Ex Post Facto Study; Quasi-Experimental Designs

**Further Readings**

Fraenkel, J. R., & Wallen, N. E. (2009). *How to design and evaluate research in education*. New York: McGraw-Hill.

Gay, L. R., Mills, G. E., & Airasian, P. (2009). *Educational research: Competencies for analysis and applications*. Upper Saddle River, NJ: Pearson Education.

Lodico, M. G., Spaulding, D. T., & Voegtle, K. H. (2006). *Methods in educational research: From theory to practice*. San Francisco: Jossey-Bass.

Mertler, C. A., & Charles, C. M. (2005). *Introduction to educational research*. Boston: Pearson.

Suter, W. N. (1998). *Primer of educational research*. Boston: Allyn and Bacon.

## Cause and Effect

Cause and effect refers to a relationship between two phenomena in which one phenomenon is the reason behind the other. For example, eating too much fast food without any physical activity leads to weight gain. Here eating without any physical activity is the "cause" and weight gain is the "effect." Another popular example in the discussion of cause and effect is that of smoking and lung cancer. A question that has surfaced in cancer research in the past several decades is, What is the effect of smoking on an individual's health? Also asked is the question, Does smoking cause lung cancer? Using data from observational studies, researchers have long established the relationship between smoking and the incidence of lung cancer; however, it took compelling evidence from several studies over several decades to establish smoking as a "cause" of lung cancer.

The term *effect* has been used frequently in scientific research. Most of the time, it can be seen that a statistically significant result from a linear regression or correlation analysis between two variables $X$ and $Y$ is explained as effect. Does $X$ really cause $Y$ or just relate to $Y$? The association (correlation) of two variables with each other in the statistical sense does not imply that one is the cause and the other is the effect. There needs to be a mechanism that explains the relationship in order for the association to be a causal one. For example, without the discovery of the substance nicotine in tobacco, it would have been difficult to establish the causal relationship between smoking and lung cancer. Tobacco companies have claimed that since there is not a single randomized controlled trial that establishes the differences in death from lung cancer between smokers and nonsmokers, there was no causal relationship. However, a cause-and-effect relationship is established by observing the same phenomenon in a wide variety of settings while controlling for other suspected mechanisms.

*Statistical correlation* (e.g., association) describes how the values of variable $Y$ of a specific population are associated with the values of another variable $X$ from the same population. For example, the death rate from lung cancer increases with increased age in the general population. The association or correlation describes the situation that there is a relationship between age and the death rate from lung cancer. *Randomized prospective studies* are often used as a tool to establish a causal effect. Time is a key element in causality because the cause must happen prior to the effect. Causes are often referred to as *treatments* or *exposures* in a study. Suppose a causal relationship between an investigational drug $A$ and response $Y$ needs to be established. Suppose $Y_A$ represents the

response when the participant is treated using $A$ and $Y_0$ is the response when the subject is treated with placebo under the same conditions. The causal effect of the investigational drug is defined as the population average $\delta = E(Y_A - Y_0)$. However, a person cannot be treated with both placebo and Treatment $A$ under the same conditions. Each participant in a randomized study will have, usually, equal potential of receiving Treatment $A$ or the placebo. The responses from the treatment group and the placebo group are collected at a specific time after exposure to the treatment or placebo. Since participants are randomized to the two groups, it is expected that the conditions (represented by covariates) are balanced between the two groups. Therefore, randomization controls for other possible causes that can affect the response $Y$, and hence the difference between the average responses from the two groups, can be thought of an estimated causal effect of treatment $A$ on $Y$.

Even though a randomized experiment is a powerful tool for establishing a causal relationship, a randomization study usually needs a lot of resources and time, and sometimes it cannot be implemented for ethical or practical reasons. Alternatively, an *observational study* may be a good tool for causal inference. In an observational study, the probability of receiving (or not receiving) treatment is assessed and accounted for. In the example of the effect of smoking on lung cancer, smoking and not smoking are the treatments. However, for ethical reasons, it is not practical to randomize subjects to treatments. Therefore, researchers had to rely on observational studies to establish the causal effect of smoking on lung cancer.

Causal inference plays a significant role in medicine, epidemiology, and social science. An issue about the average treatment effect is also worth mentioning. The average treatment effect, $\delta = E(Y_1) - E(Y_2)$, between two treatments is defined as the difference between two outcomes, but, as mentioned previously, a subject can receive only one of "rival" treatments. In other words, it is impossible for a subject to have two outcomes at the same time. $Y_1$ and $Y_2$ are called *counterfactual outcomes*. Therefore, the average treatment effect can never be observed. In the causal inference literature, several estimating methods of average treatment effect are proposed to deal with this obstacle. Also, for observational study, estimators for average treatment effect with confounders controlled have been proposed.

*Abdus S. Wahed and Yen-Chih Hsu*

***See also*** Clinical Trial; Observational Research; Randomization Tests

**Further Readings**

Freedman, D. (2005). *Statistical models: Theory and practice*. Cambridge, UK: Cambridge University Press.
Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*(396), 945–960.

# Ceiling Effect

The term *ceiling effect* is a measurement limitation that occurs when the highest possible score or close to the highest score on a test or measurement instrument is reached, thereby decreasing the likelihood that the testing instrument has accurately measured the intended domain. A ceiling effect can occur with questionnaires, standardized tests, or other measurements used in research studies. A person's reaching the ceiling or scoring positively on all or nearly all the items on a measurement instrument leaves few items to indicate whether the person's true level of functioning has been accurately measured. Therefore, whether a large percentage of individuals reach the ceiling on an instrument or whether an individual scores very high on an instrument, the researcher or interpreter has to consider that what has been measured may be more of a reflection of the parameters of what the instrument is able to measure than of how the individuals may be ultimately functioning. In addition, when the upper limits of a measure are reached, discriminating between the functioning of individuals within the upper range is difficult. This entry focuses on the impact of ceiling effects on the interpretation of research results, especially the results of standardized tests.

## Interpretation of Research Results

When a ceiling effect occurs, the interpretation of the results attained is impacted. For example,

a health survey may include a range of questions that focus on the low to moderate end of physical functioning (e.g., individual is able to walk up a flight of stairs without difficulty) versus a range of questions that focus on higher levels of physical functioning (e.g., individual is able to walk at a brisk pace for 1 mile without difficulty). Questions within the range of low to moderate physical functioning provide valid items for individuals on that end of the physical functioning spectrum rather than for those on the higher end of the physical functioning spectrum. Therefore, if an instrument geared toward low to moderate physical functioning is administered to individuals with physical health on the upper end of the spectrum, a ceiling effect will likely be reached in a large portion of the cases, and interpretation of their ultimate physical functioning would be limited.

A ceiling effect can be present within results of a research study. For example, a researcher may administer the health survey described in the previous paragraph to a treatment group in order to measure the impact of a treatment on overall physical health. If the treatment group represents the general population, the results may show a large portion of the treatment group to have benefited from the treatment because they have scored high on the measure. However, this high score may signify the presence of a ceiling effect, which calls for caution when one is interpreting the significance of the positive results. If a ceiling effect is suspected, an alternative would be to use another measure that provides items that target better physical functioning. This would allow participants to demonstrate a larger degree of differentiation in physical functioning and provide a measure that is more sensitive to change or growth from the treatment.

## Standardized Tests

The impact of the ceiling effect is important when interpreting standardized test results. Standardized tests have higher rates of score reliability because the tests have been administered to a large sampling of the population. The large sampling provides a scale of standard scores that reliably and validly indicates how close a person who takes the same test performs compared with the mean performance on the original sampling. Standardized tests include aptitude tests such as the Wechsler Intelligence Scales, the Stanford-Binet Scales, and the Scholastic Aptitude Test, and achievement tests such as the Iowa Test of Basic Skills and the Woodcock-Johnson: Tests of Achievement.

When individuals score at the upper end of a standardized test, especially 3 standard deviations from the mean, then the ceiling effect is a factor. It is reasonable to conclude that such a person has exceptional abilities compared with the average person within the sampled population, but the high score is not necessarily a highly reliable measure of the person's true ability. What may have been measured is more a reflection of the test than of the person's true ability. In order to attain an indication of the person's true ability when the ceiling has been reached, an additional measure with an increased range of difficult items would be appropriate to administer. If such a test is not available, the test performance would be interpreted stating these limitations.

The ceiling effect should also be considered when one is administering a standardized test to an individual who is at the top end of the age range for a test and who has elevated skills. In this situation, the likelihood of a ceiling effect is high. Therefore, if a test administrator is able to use a test that places the individual on the lower age end of a similar or companion test, the chance of a ceiling effect would most likely be eliminated. For example, the Wechsler Intelligence Scales have separate measures to allow for measurement of young children, children and adolescents, and older adolescents and adults. The upper end and lower end of the measures overlap, meaning that a 6- to 7-year-old could be administered the Wechsler Preschool and Primary Scale of Intelligence or the Wechsler Intelligence Scale for Children. In the event a 6-year-old is cognitively advanced, the Wechsler Intelligence Scale for Children would be a better choice in order to avoid a ceiling effect.

It is important for test developers to monitor ceiling effects on standardized instruments as they are utilized in the public. If a ceiling effect is noticed within areas of a test over time, those elements of the measure should be improved to provide better discrimination for high performers. The rate of individuals scoring at the upper end

of a measure should coincide with the standard scores and percentiles on the normal curve.

*Tish Holub Taylor*

*See also* Instrumentation; Standardized Score; Survey; Validity of Measurement

**Further Readings**

Austin, P. C., & Bruner, L. J. (2003). Type I error inflation in the presence of a ceiling effect. *American Statistician, 57*(2), 97–105.

Gunst, R. F., & Barry, T. E. (2003). One way to moderate ceiling effects. *Quality Progress, 36*(10), 84–86.

Kaplan, C. (1992). Ceiling effects in assessing high-IQ children with the WPPSI-R. *Journal of Clinical Child Psychology, 21*(4), 403–406.

Rifkin, B. (2005). A ceiling effect in traditional classroom foreign language instruction: Data from Russian. *Modern Language Journal, 89*(1), 3–18.

Sattler, J. M. (2001). *Assessment of children: Cognitive applications* (4th ed.). San Diego, CA: Jerome M. Sattler.

Taylor, R. L. (1997). *Assessment of exceptional students: Educational and psychological procedures* (4th ed.). Boston: Allyn and Bacon.

Uttl, B. (2005). Measurement of individual differences: Lessons from memory assessment in research and clinical practice. *Psychological Science, 16*(6), 460–467.

## Central Limit Theorem

The central limit theorem (CLT) is, along with the theorems known as *laws of large numbers,* the cornerstone of *probability theory.* In simple terms, the theorem describes the distribution of the sum of a large number of random numbers, all drawn independently from the same probability distribution. It predicts that, regardless of this distribution, as long as it has finite variance, then the sum follows a precise law, or distribution, known as the *normal distribution.*

Let us describe the normal distribution with mean $\mu$ and variance $\sigma^2$: It is defined through its density function,

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2},$$

where the variable $x$ ranges from $-\infty$ to $+\infty$. This means that if a random variable follows this distribution, then the probability that it is larger than $a$ and smaller than $b$ is equal to the integral of the function $f(x)$ (the area under the graph of the function) from $x = a$ to $x = b$. The normal density is also known as Gaussian density, named for Carl Friedrich Gauss, who used this function to describe astronomical data. If we put $\mu = 0$ and $\sigma^2 = 1$ in the above formula, then we obtain the so-called *standard normal density.*

In precise mathematical language, the CLT states the following: Suppose that $X_1, X_2, \ldots$ are independent random variables with the same distribution, having *mean* $\mu$ and *variance* $\sigma^2$ but being otherwise arbitrary. Let $S_n = X_1 + \cdots + X_n$ be their sum. Then

$$P(a < \frac{S_n - n\mu}{\sqrt{n\sigma^2}} < b) \to \int_a^b \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt, \quad \text{as } n \to \infty.$$

It is more appropriate to define the standard normal density as the density of a random variable $\zeta$ with zero mean and variance 1 with the property that, for every $a$ and $b$ there is $c$ such that if $\zeta_1, \zeta_2$ are independent copies of $\zeta$, then $a\zeta_1 + b\zeta_2$ is a copy of $c\zeta$ . It follows that $a^2 + b^2 = c^2$ holds and that there is only one choice for the density of $\zeta$, namely, the standard normal density.

As an example, consider tossing a fair coin $n = 1,000$ times and determining the probability that fewer than 450 heads are obtained. The CLT can be used to give a good approximation of this probability. Indeed, if we let $X_i$ be a random variable that takes value 1 if heads show up at the $i$th toss or value 0 if tails show up, then we see that the assumptions of the CLT are satisfied because the random variables have the same mean $\mu = 1/2$ and variance $\sigma^2 = 1/4$. On the other hand, $S_n = X_1 + \cdots + X_n$ is the number of heads. Since $S_n \leq 450$ if and only if $(S_n - n\mu)/\sqrt{n\sigma^2} \leq (450 - 500)/\sqrt{250} = -3.162$, we find, by the CLT, that the probability that we get at most 450 heads equals the integral of the standard density from $-\infty$ to $-3.162$. This integral can be computed with the help of a computer (or tables in olden times) and found to be about 0.00078, which is a reasonable approximation. Incidentally, this kind of thing leads to the so-called *statistical hypothesis testing:* If we

toss a coin and see 430 heads and 570 tails, then we should be suspicious that the coin is not fair.

## Origins

The origins of the CLT can be traced to a paper by Abraham de Moivre (1733), who described the CLT for symmetric Bernoulli trials; that is, in tossing a fair coin $n$ times, the number $S_n$ of heads has a distribution that is approximately that of $\frac{n}{2} + \frac{\sqrt{n}}{2}\zeta$. The result fell into obscurity but was revived in 1812 by Pierre-Simon Laplace, who proved and generalized de Moivre's result to asymmetric Bernoulli trials (weighted coins). Nowadays, this particular case is known as the *de Moivre–Laplace CLT* and is usually proved using *Stirling's approximation* for the product $n!$ of the first $n$ positive integers: $n! \approx n^n e^{-n}\sqrt{2\pi n}$. The factor $\sqrt{2\pi}$ here is the same as the one appearing in the normal density. Two 19th-century Russian mathematicians, Pafnuty Chebyshev and A. A. Markov, generalized the CLT of their French predecessors and proved it using the *method of moments*.

## The Modern CLT

The modern formulation of the CLT gets rid of the assumption that the summands be identically distributed random variables. Its most general and useful form is that of the Finnish mathematician J. W. Lindeberg (1922). It is stated for *triangular arrays,* that is, random variables $S_n = X_{n,1}, X_{n,2}, \ldots, X_{n,k_n}$, depending on two indices that, for fixed $n$, are independent with respect to the second index. Letting $S_n = X_1 + \cdots + X_{n,k_n}$ be the sum with respect to the second index, we have that $S_n$ minus its mean, $ES_n$, and divided by its standard deviation, has a distribution that, as $n$ tends to $\infty$, is standard normal, provided an asymptotic negligibility condition for the variances holds. A version of Lindeberg's theorem was formulated by A. Liapounoff in 1901. It is worth pointing out that Liapounoff introduced a new proof technique based on *characteristic functions* (also known as *Fourier transforms*), whereas Lindeberg's technique was an ingenious step-by-step replacement of the general summands by normal ones. Liapounoff's theorem has a condition that is weaker than that of Lindeberg but is quite good in practice. In 1935, Paul Lévy and William Feller established necessary conditions for the validity of the CLT. In 1951, H. F. Trotter gave an elementary analytical proof of the CLT.

## The Functional CLT

The functional CLT is stated for summands that take values in a multidimensional (one talks of random vectors in a Euclidean space) or even infinite-dimensional space. A very important instance of the functional CLT concerns convergence to *Brownian motion,* which also provides a means of defining this most fundamental object of modern probability theory. Suppose that $S_n$ is as stated in the beginning of this entry. Define a function $s_n(t)$ of "time" as follows: At time $t = k/n$, where $k$ is a positive integer, let $s_n(t)$ have value $(S_k - \mu k)/\sigma\sqrt{n}$; now join the points $[k/n, s_n(k/n)]$, $[(k+1)/n, s_n(k+1)/n)]$ by a straight line segment, for each value of $k$, to obtain the graph of a continuous random function $s_n(t)$. *Donsker's theorem* states that the probability distribution of the random function $s_n(t)$ converges (in a certain sense) to the distribution of a random function, or stochastic process, which can be defined to be the standard Brownian motion.

## Other Versions and Consequences of the CLT

Versions of the CLT for *dependent random variables* also exist and are very useful in practice when independence is either violated or not possible to establish. Such versions exist for Markov processes, for regenerative processes, and for martingales, as well as others.

The work on the CLT gave rise to the general area known as *weak convergence (of stochastic processes),* the origin of which is in Yuri Prokhorov (1956) and Lucien Le Cam (1957). Nowadays, every result of CLT can (and should) be seen in the light of this general theory.

Finally, it should be mentioned that all flavors of the CLT rely heavily on the assumption that the probability that the summands be very large is small, which is often cast in terms of finiteness of a moment such as the variance. In a variety of applications, *heavy-tailed random variables* do not satisfy this assumption. The study of limit theorems of sums of independent heavy-tailed

random variables leads to the so-called *stable distributions,* which are generalizations of the normal law, and like the normal law, they are "preserved by linear transformations," but they all have infinite variance. Boris Gnedenko and A. N. Kolmogorov call such limit results central limit theorems as well, but the terminology is not well established, even though the research, applications, and use of such results are quite extended nowadays. A functional CLT (an analog of Donsker's theorem) for sums of heavy-tailed random variables gives rise to the so-called *stable processes,* which are special cases of processes with stationary and independent increments known as *Lévy processes.*

## Uses of the CLT

The uses of the CLT are numerous. First, it is a cornerstone of *statistics.* Indeed, one could argue that without the CLT, there would be no such subject; for example, confidence intervals and parameter and density estimation make explicit use of the CLT or a version of it known as the *local limit theorem,* which, roughly speaking, states that the probability density function of $S_n$ converges to the normal density. Of particular importance are dependent versions of the CLT, which find applications in the modern areas of data mining and artificial intelligence.

In connection with statistics, of importance is the question of speed of convergence to CLT. Answers to this are given by the classical Berry-Esséen theorem or by more modern methods consisting of embedding the random sums into a Brownian motion (via the so-called Hungarian construction of Komlós, Major, and Tusnády).

In *applied probability,* various stochastic models depending on normal random variables or Brownian motions can be justified via (functional) central limit theorems. Areas within applied probability benefiting from the use of the CLT are, among others, stochastic networks, particle systems, stochastic geometry, the theory of risk, mathematical finance, and stochastic simulation.

There are also various applications of the CLT in engineering, physics (especially in statistical mechanics), and mathematical biology.

### *Example 1: Confidence Intervals.*

Suppose we set up an experiment in which we try to find out how the glucose blood level in a certain population affects a disease. We introduce a mathematical model as follows: Assume that the glucose level data we collect are observations of independent random variables, each following a probability distribution with an unknown mean $\theta$ (the "true" glucose level) and wish to estimate this $\theta$. Letting $X_1, \ldots, X_n$ be the data we have, a reasonable estimate for $\theta$ is the sample mean $\overline{X} = S_n/n = (X_1 + \cdots + X_n)/n$. A $p$-confidence interval (where $p$ is a large probability, say $p = .95$) is an interval such that the probability that it contains $\overline{X}$ is at least $p$. To find this confidence interval, we use the central limit theorem to justify that $n(\overline{X} - \frac{\sigma}{\sqrt{n}}) = (S_n - n\theta)/\sigma\sqrt{n}$ is approximately standard normal. We then find the number $\alpha$ such that $P(-\alpha < \zeta < \alpha) = p$, where $\zeta$ is a standard normal random variable, and, on replacing $\zeta$ by $n(\overline{X} - \frac{\sigma}{\sqrt{n}})$, we have that $P(\overline{X} - \alpha\sigma\sqrt{n} < \theta < \overline{X} + \alpha\sigma\sqrt{n}) = p$, meaning that the interval centered at $\overline{X}$ and having width $2\alpha\sigma\sqrt{n}$ is a $p$-confidence interval. One way to deal with the fact that $\sigma$ may be unknown is to replace it by the sample standard deviation.

### *Example 2: An Invariance Principle*

A very important, for applications, consequence of the CLT (or, more generally, of the principle of weak convergence mentioned above) is that it yields information about functions of the approximating sequence. For example, consider a fair game of chance, say tossing a fair coin whereby \$1 is won on appearance of heads or lost on appearance of tails. Let $S_n$ be the net winnings in $n$ coin tosses (negative if there is a net loss). Define the quantity $T_n$ roughly as the number of times before $n$ that the gambler had a net loss. By observing that this quantity is a function of the function $s_n(t)$ discussed earlier, and using the Donsker's CLT, it can be seen that the distribution function of $T_n/n$ has, approximately, the arcsine law: $P(T_n/n \leq x) \approx (2/\pi)\arcsin\sqrt{x}$, for large $n$.

*Takis Konstantopoulos*

***See also*** Distribution; Law of Large Numbers

**Further Readings**

Billingsley, P. (1968). *Convergence of probability measures*. New York: Wiley.
de Moivre, A. (1738). *The doctrine of chances* (2nd ed.). London: Woodfall.
de Moivre, A. (1926). Approximatio ad summam terminorum binomii $\frac{n}{a+b}$ in seriem expansi [Approximating the sum of the terms of the binomial (a + b)n expanded into a series]. Reprinted by R. C. Archibald, A rare pamphlet of de Moivre and some of his discoveries. *Isis, 8,* 671–684. (Original work published 1733)
Donsker, M. (1951). An invariance principle for certain probability limit theorems. *Memoirs of the American Mathematical Society, 6,* 1–10.
Feller, W. (1935). Über den Zentralen Grenzwertsatz des Wahrscheinlichkeitsrechnung [Central limit theorem of probability theory]. *Math. Zeitung, 40,* 521–559.
Feller, W. (1937). Über den Zentralen Grenzwertsatz des Wahrscheinlichkeitsrechnung [Central limit theorem of probability theory]: Part II. *Math. Zeitung, 42,* 301–312.
Gnedenko, B. V., & Kolmogorov, A. N. (1968). *Limit theorems for sums of independent random variables* (K. L. Chung, Trans.). Reading, MA: Addison-Wesley. (Original work published 1949)
Komlós, J., Major, P., & Tusnády, G. (1975). An approximation of partial sums of independent RV'-s, and the sample DF. *Zeit. Wahrscheinlichkeitstheorie Verw.* 32, 111–131.
Laplace, P.-S. (1992). *Théorie analytique des probabilités* [Analytical theory of probability]. Paris: Jacques Gabay. (Original work published 1828)
Le Cam, L. (1957). Convergence in distribution of stochastic processes. *University of California Publications in Statistics, 2,* 207–236.
Lévy, P. (1935). Propriétés asymptotiques des sommes de variables aléatoires indépendantes ou enchaînées [Asymptotic properties of sums of independent or dependent random variables]. *Journal des Mathématiques Pures et Appliquées, 14,* 347–402.
Liapounoff, A. (1901). Nouvelle forme du théorème sur la limite de probabilités [New form of a theory on probability limits]. *Memoirs de l'Académie Imperial des Sciences de Saint-Pétersbourg, 13,* 359–386.
Lindeberg, J. W. (1922). Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung [A new investitgation into the theory of calculating probabilities]. *Mathematische Zeitschrift, 15,* 211–225.
Prokhorov, Y. (1956). Convergence of random processes and limit theorems in probability theory. *Theory of Probability and Its Applications, 1,* 157–214.
Trotter, H. F. (1959). Elementary proof of the central limit theorem. *Archiv der Mathematik, 10,* 226–234.
Whitt, W. (2002). *Stochastic-process limits*. New York: Springer-Verlag.

# Central Tendency, Measures of

One of the most common statistical analyses used in descriptive statistics is a process to determine where the average of a set of values falls. There are multiple ways to determine the middle of a group of numbers, and the method used to find the average will determine what information is known and how that average should be interpreted. Depending on the data one has, some methods for finding the average may be more appropriate than others.

The *average* describes the typical or most common number in a group of numbers. It is the one value that best represents the entire group of values. Averages are used in most statistical analyses, and even in everyday life. If one wanted to find the typical house price, family size, or score on a test, some form of average would be computed each time. In fact, one would compute a different type of average for each of those three examples.

Researchers use different ways to calculate the average, based on the types of numbers they are examining. Some numbers, measured on the nominal level of measurement, are not appropriate to do some types of averaging. For example, if one were examining the variable of types of vegetables, and the labels of the levels were cucumbers, zucchini, carrots, and turnips, if one performed some types of average, one may find that the average vegetable was 3/4 carrot and 1/4 turnip, which makes no sense at all. Similarly, using some types of average on interval-level continuous variables may result in an average that is very imprecise and not very representative of the sample one is using. When the three main methods for examining the average are collectively discussed, they are referred to as *measures of central tendency.* The three primary measures of central tendency commonly used by researchers are the *mean,* the *median,* and the *mode.*

## Mean

The mean is the most commonly used (and misused) measure of central tendency. The mean is defined as the sum of all the scores in the sample, divided by the number of scores in the sample. This type of mean is also referred to as the arithmetic mean, to distinguish it from other types of means, such as the geometric mean or the harmonic mean. Several common symbols or statistical notations are used to represent the mean, including $\overline{x}$, which is read as *x-bar* (the mean of the sample), and $\mu$, which is read as *mu* (the mean of the population). Some research articles also use an italicized uppercase letter *M* to indicate the sample mean.

Much as different symbols are used to represent the mean, different formulas are used to calculate the mean. The difference between the two most common formulas is found only in the symbols used, as the formula for calculating the mean of a sample uses the symbols appropriate for a sample, and the other formula is used to calculate the mean of a population, and as such uses the symbols that refer to a population.

For calculating the mean of a sample, use

$$\overline{x} = \Sigma x/n.$$

For calculating the mean of a population, use

$$\mu = \Sigma X/N.$$

Calculating the mean is a very simple process. For example, if a student had turned in five homework assignments that were worth 10 points each, the student's scores on those assignments might have been 10, 8, 5, 7, and 9. To calculate the student's mean score on the homework assignments, first one would add the values of all the homework assignments:

$$10 + 8 + 5 + 7 + 9 = 39.$$

Then one would divide the sum by the total number of assignments (which was 5):

$$39/5 = 7.8.$$

For this example, 7.8 would be the student's mean score for the five homework assignments. That indicates that, on average, the student scored a 7.8 on each homework assignment.

### *Other Definitions*

Other definitions and explanations can also be used when interpreting the mean. One common description is that the mean is like a balance point. The mean is located at the center of the values of the sample or population. If one were to line up the values of each score on a number line, the mean would fall at the exact point where the values are equal on each side; the mean is the point closest to the squared distances of all the scores in the distribution.

Another way of thinking about the mean is as the amount per individual, or how much each individual would receive if one were to divide the total amount equally. If Steve has a total of \$50, and if he were to divide it equally among his four friends, each friend would receive \$50/4 = \$12.50. Therefore, \$12.50 would be the mean.

The mean has several important properties that are found only with this measure of central tendency. If one were to change a score in the sample from one value to another value, the calculated value of the mean would change. The value of the mean would change because the value of the sum of all the scores would change, thus changing the numerator. For example, given the scores in the earlier homework assignment example (10, 8, 5, 7, and 9), the student previously scored a mean of 7.8. However, if one were to change the value of the score of 9 to a 4 instead, the value of the mean would change:

$$10 + 8 + 5 + 7 + 4 = 34$$
$$34/5 = 6.8.$$

By changing the value of one number in the sample, the value of the mean was lowered by one point. Any change in the value of a score will result in a change in the value of the mean.

If one were to remove or add a number to the sample, the value of the mean would also change, as then there would be fewer (or greater) numbers in the sample, thus changing the numerator and the denominator. For example, if one were to calculate the mean of only the first four homework assignments (thereby removing a number from the sample),

$$10 + 8 + 5 + 7 = 30$$
$$30/4 = 7.5.$$

If one were to include a sixth homework assignment (adding a number to the sample) on which the student scored an 8,

$$10 + 8 + 5 + 7 + 9 + 8 = 47$$
$$47/6 = 7.83.$$

Either way, whether one adds or removes a number from the sample, the mean will almost always change in value. The only instance in which the mean will not change is if the number that is added or removed is exactly equal to the mean. For example, if the score on the sixth homework assignment had been 7.8,

$$10 + 8 + 5 + 7 + 9 + 7.8 = 46.8$$
$$46.8/6 = 7.8.$$

If one were to add (or subtract) a constant to each score in the sample, the mean will increase (or decrease) by the same constant value. So if the professor added three points to the score on every homework assignment,

$$(10 + 3) + (8 + 3) + (5 + 3) + (7 + 3) + (9 + 3)$$
$$13 + 11 + 8 + 10 + 12 = 54$$
$$54/5 = 10.8.$$

The mean homework assignment score increased by 3 points. Similarly, if the professor took away two points from each original homework assignment score,

$$(10 - 2) + (8 - 2) + (5 - 2) + (7 - 2) + (9 - 2)$$
$$8 + 6 + 3 + 5 + 7 = 29;$$
$$29/5 = 5.8.$$

The mean homework assignment score decreased by two points. The same type of situation will occur with multiplication and division. If one multiplies (or divides) every score by the same number, the mean will also be multiplied (or divided) by that number. Multiplying the five original homework scores by four,

$$10(4) + 8(4) + 5(4) + 7(4) + 9(4)$$
$$40 + 32 + 20 + 28 + 36 = 156$$
$$156/5 = 31.2,$$

results in the mean being multiplied by four as well. Dividing each score by 3,

$$10/3 + 8/3 + 5/3 + 7/3 + 9/3$$
$$3.33 + 2.67 + 1.67 + 2.33 + 3 = 13$$
$$13/5 = 2.6,$$

will result in the mean being divided by three as well.

### *Weighted Mean*

Occasionally, one will need to calculate the mean of two or more groups, each of which has its own mean. In order to get the overall mean (sometimes called the grand mean or weighted mean), one will need to use a slightly different formula from the one used to calculate the mean for only one group:

$$\text{Weighted mean} = (\Sigma x_1 + \Sigma x_2 + \ldots \Sigma x_n)/(n_1 + n_2 + \ldots n_n).$$

To calculate the weighted mean, one will divide the sum of all the scores in every group by the number of scores in every group. For example, if Carla taught three classes, and she gave each class a test, the first class of 25 students might have a mean of 75, the second class of 20 students might have a mean of 85, and the third class of 30 students might have a mean of 70. To calculate the weighted mean, Carla would first calculate the summed scores for each class, then add the summed scores together, then divide by the total number of students in all three classes. To find the summed scores, Carla will need to rework the formula for the sample mean to find the summed scores instead:

$$\overline{X} = \Sigma X/n \text{ becomes } \Sigma x = \overline{x}(n).$$

With this reworked formula, find each summed score first:

$$\Sigma x_1 = 75(25) = 1,875$$
$$\Sigma X_2 = 85(20) = 1,700$$
$$\Sigma X_3 = 70(30) = 2,100.$$

Next, add the summed scores together:

$$1,875 + 1,700 + 2,100 = 5,675.$$

Then add the number of students per class together:

$$25 + 20 + 30 = 75.$$

Finally, divide the total summed scores by the total number of students:

$$5,675/75 = 75.67$$

This is the weighted mean for the test scores of the three classes taught by Carla.

## Median

The median is the second measure of central tendency. It is defined as the score that cuts the distribution exactly in half. Much as the mean can be described as the balance point, where the values on each side are identical, the median is the point where the number of scores on each side is equal. As such, the median is influenced more by the number of scores in the distribution than by the values of the scores in the distribution. The median is also the same as the 50th percentile of any distribution. Generally the median is not abbreviated or symbolized, but occasionally *Mdn* is used.

The median is simple to identify. The method used to calculate the median is the same for both samples and populations. It requires only two steps to calculate the median. In the first step, order the numbers in the sample from lowest to highest. So if one were to use the homework scores from the mean example, 10, 8, 5, 7, and 9, one would first order them 5, 7, 8, 9, 10. In the second step, find the middle score. In this case, there is an odd number of scores, and the score in the middle is 8.

5 7 **8** 9 10

Notice that the median that was calculated is not the same as the mean for the same sample of homework assignments. This is because of the different ways in which those two measures of central tendency are calculated.

It is simple to find the middle when there are an odd number of scores, but it is a bit more complex when the sample has an even number of scores. For example, when there were six homework scores (10, 8, 5, 7, 9, and 8), one would still line up the homework scores from lowest to highest, then find the middle:

5 7 8 | 8 9 10.

In this example, the median falls between two identical scores, so one can still say that the median is 8. If the two middle numbers were different, one would find the middle number between the two numbers. For example, if one increased one of the student's homework scores from an 8 to a 9,

5 7 8 | 9 9 10.

In this case, the middle falls halfway between 8 and 9, at a score of 8.5.

Statisticians disagree over the correct method for calculating the median when the distribution has multiple repeated scores in the center of the distribution. Some statisticians use the methods described above to find the median, whereas others believe the scores in the middle need to be reduced to fractions to find the exact midpoint of the distribution. So in a distribution with the following scores,

2 3 3 4 5 5 5 5 5 6,

some statisticians would say the median is 5, whereas others (using the fraction method) would report the median as 4.7.

## Mode

The mode is the last measure of central tendency. It is the value that occurs most frequently. It is the simplest and least precise measure of central tendency. Generally, in writing about the mode, scholars label it simply *mode,* although some books or papers use *Mo* as an abbreviation. The method for finding the mode is the same for both samples and populations. Although there are several ways one could find the mode, a simple method is to list each score that appears in the

sample. The score that appears the most often is the mode. For example, given the following sample of numbers,

$$3, 4, 6, 2, 7, 4, 5, 3, 4, 7, 4, 2, 6, 4, 3, 5,$$

one could arrange them in numerical order:

$$2,\ 2,\ 3,\ 3,\ 3,\ 4,\ 4,\ 4,\ 4,\ 4,\ 5,\ 5,\ 6,\ 6,\ 7,\ 7.$$

Once the numbers are arranged, it becomes apparent that the most frequently appearing number is 4:

$$2, 2, 3, 3, 3, \mathbf{4}, \mathbf{4}, \mathbf{4}, \mathbf{4}, \mathbf{4}, 5, 5, 6, 6, 7, 7.$$

Thus 4 is the mode of that sample of numbers. Unlike the mean and the median, it is possible to have more than one mode. If one were to add two threes to the sample,

$$2,\ 2,\ 3,\ 3,\ 3,\ 3,\ 3,\ 4, 4,\ 4,\ 4,\ 4,\ 5,\ 5,$$
$$6,\ 6,\ 7,\ 7,$$

then both 3 and 4 would be the most commonly occurring number, and the mode of the sample would be 3 and 4. The term used to describe a sample with two modes is *bimodal.* If there are more than two modes in a sample, one says the sample is *multimodal.*

## When to Use Each Measure

Because each measure of central tendency is calculated with a different method, each measure is different in its precision of measuring the middle, as well as which numbers it is best suited for.

### *Mean*

The mean is often used as the default measure of central tendency. As most people understand the concept of average, they tend to use the mean whenever a measure of central tendency is needed, including times when it is not appropriate to use the mean. Many statisticians would argue that the mean should not be calculated for numbers that are measured at the nominal or ordinal levels of measurement, due to difficulty in the interpretation of the results. The mean is also used in other statistical analyses, such as calculations of standard deviation. If the numbers in the sample are fairly normally distributed, and there is no specific reason that one would want to use a different measure of central tendency, then the mean should be the best measure to use.

### *Median*

There are several reasons to use the median instead of the mean. Many statisticians believe that it is inappropriate to use the mean to measure central tendency if the distribution was measured at the ordinal level. Because variables measured at the ordinal level contain information about direction but not distance, and because the mean is measured in terms of distance, using the mean to calculate central tendency would provide information that is difficult to interpret.

Another occasion to use the median is when the distribution contains an *outlier.* An outlier is a value that is very different from the other values. Outliers tend to be located at the far extreme of the distribution, either high or low. As the mean is so sensitive to the value of the scores, using the mean as a measure of central tendency in a distribution with an outlier would result in a nonrepresentative score. For example, looking again at the five homework assignment scores, if one were to replace the nine with a score of 30,

$$5\ 7\ 8\ 9\ 10 \text{ becomes } 5\ 7\ 8\ 30\ 10$$
$$5 + 7 + 8 + 30 + 10 = 60$$
$$60/5 = 12.$$

By replacing just one value with an outlier, the newly calculated mean is not a good representation of the average values of our distribution. The same would occur if one replaced the score of 9 with a very low number:

$$5\ 7\ 8\ 9\ 10 \text{ becomes } 5\ 7\ 8\ -4\ 10$$
$$5 + 7 + 8 + -4 + 10 = 26$$
$$26/5 = 5.2.$$

Since the mean is so sensitive to outliers, it is best to use the median for calculating central tendency. Examining the previous example, but using the median,

$$5\ \ 7\ \ 8\ \ -4\ \ 10$$
$$-4\ \ 5\ \ 7\ \ \ 8\ \ \ 10.$$

The middle number is 7, therefore the median is 7, a much more representative number than the mean of that sample, 5.2.

If the numbers in the distribution were measured on an item that had an open-ended option, one should use the median as the measure of central tendency. For example, a question that asks for demographic information such as age or salary may include either a lower or upper category that is open ended:

| *Number of Cars Owned* | *Frequency* |
|---|---|
| 0 | 4 |
| 1 | 10 |
| 2 | 16 |
| 3 or more | 3 |

The last answer option is open ended because an individual with three cars would be in the same category as someone who owned 50 cars. As such, it is impossible to accurately calculate the mean number of cars owned. However, it is possible to calculate the median response. For the above example, the median would be 2 cars owned.

A final condition in which one should use the median instead of the mean is when one has incomplete information. If one were collecting survey information, and some of the participants refused or forgot to answer a question, one would have responses from some participants but not others. It would not be possible to calculate the mean in this instance, as one is missing important information that could change the mean that would be calculated if the missing information were known.

### *Mode*

The mode is well suited to be used to measure the central tendency of variables measured at the nominal level. Because variables measured at the nominal level are given labels, then any number assigned to these variables does not measure quantity. As such, it would be inappropriate to use the mean or the median with variables measured at this level, as both of those measures of central tendency require calculations involving quantity. The mode should also be used when finding the middle of distribution of a discrete variable. As these variables exist only in whole numbers, using other methods of central tendency may result in fractions of numbers, making it difficult to interpret the results. A common example of this is the saying, "The average family has a mom, a dad, and 2.5 kids." People are discrete variables, and as such, they should never be measured in such a way as to obtain decimal results. The mode can also be used to provide additional information, along with other calculations of central tendency. Information about the location of the mode compared with the mean can help determine whether the distribution they are both calculated from is skewed.

*Carol A. Carman*

***See also*** Descriptive Statistics; Levels of Measurement; Mean; Median; Mode; "On the Theory of Scales of Measurement"; Results Section; Sensitivity; Standard Deviation; Variability, Measure of

**Further Readings**

Coladarci, T., Cobb, C. D., Minium, E. W., & Clarke, R. C. (2004). *Fundamentals of statistical reasoning in education*. Hoboken, NJ: Wiley.

Gravetter, F. J., & Wallnau, L. B. (2004). *Statistics for the behavioral sciences* (6th ed.). Belmont, CA: Thomson Wadsworth.

Salkind, N. J. (2008). *Statistics for people who (think they) hate statistics*. Thousand Oaks, CA: Sage.

Thorndike, R. M. (2005). *Measurement and evaluation in psychology and education*. (7th ed.). Upper Saddle River, NJ: Pearson Education.

## Change Scores

The measurement of change is fundamental in the social and behavioral sciences. Many researchers have used change scores to measure gain in ability or shift in attitude over time, or *difference scores* between two variables to measure a construct (e.g., self-concept vs. ideal self). This entry introduces estimation of change scores, its assumptions and applications, and at the end offers a recommendation on the use of change scores.

Let $Y$ and $X$ stand for the measures obtained by applying the same test to the subjects on two occasions. *Observed* change or difference score is $D = Y - X$. The *true* change is $D_T = Y_T - X_T$, where $Y_T$ and $X_T$ represent the subject's true status

at these times. The development of measuring the true change $D_T$ follows two paths, one using change score and the other using *residual change score*.

The only assumption to calculate change score is that $Y$ (e.g., posttest scores) and $X$ (e.g., pretest scores) should be on the same numerical scale; that is, the scores on posttest are comparable to scores on pretest. This only requirement does not suggest that pretest and posttest measure the same construct. Thus, such change scores can be extended to any kind of difference score between two measures (measuring the same construct or not) that are on the same numerical scale. The two measures are *linked*, as if the two scores are obtained from a single test or two observations are made by the same observer. The correlation between linked observations (e.g., two observations made by the same observer) will be higher than that between independent observations (e.g., two observations made by different observers). Such linkage must be considered in defining the *reliability coefficient* for difference scores. The reliability of change or difference scores is defined as the correlation of the scores with independently observed difference scores. The reliability for change scores produced by comparing two independent measures will most likely be smaller than that for the linked case.

Raw change or difference scores are computed with two observed measures ($D = Y - X$). Observed scores are systematically related to random error of measurement and thus unreliable. Conclusions based on these scores tend to be fallacious.

True change score is measured as the difference between the person's true status at posttest and pretest times, $D_T = Y_T - X_T$. The key is to remove the measurement error from the two observed measures. There are different ways to correct the errors in the two raw measures used to obtain raw gain scores. The first way is to correct the error in pretest scores using the reliability coefficient of $X$ and simple regression. The second way is to correct errors in both pretest and posttest scores using the reliability coefficient of both $X$ and $Y$ and simple regression. The third way is the *Lord procedure*. With this procedure, the estimates of $Y_T$ and $X_T$ are obtained by the use of a multiple regression procedure that incorporates the reliability of a measure (e.g., $X$) and information that can be borrowed from the other measure (e.g., $Y$). The estimator obtained with the Lord procedure is better than those from the previous two ways and the raw change score in that it yields a smaller mean square of deviation between the estimate and the true change ($\hat{D}_T - D_T$).

Lee Cronbach and Lita Furby, skeptical of change scores, proposed a better estimate of true change by incorporating information in regression of two additional categories of variables besides the pretest and posttest measures used in Lord's procedure. The two categories of variables include Time 1 measures $W$ (e.g., experience prior to the treatment, demographic variables, different treatment group membership) and Time 2 measures $Z$ (e.g., a follow-up test a month after the treatment). $W$ and $X$ need not be simultaneous, nor do $Z$ and $Y$. Note that $W$ and $Z$ can be multivariate. Additional information introduced by $W$ and $Z$ improves the estimation of true change with smaller mean squares of ($\hat{D}_T - D_T$) if the sample size is large enough for the weight of the information related to $W$ and $Z$ to be accurately estimated.

The other approach to estimating the true change $D_T$ is to use the residual change score. The development of the residual change score estimate is similar to that of the change score estimate. The raw residual-gain score is obtained by regressing the observed posttest measure $Y$ on the observed pretest measure $X$. In this way, the portion in the posttest score that can be predicted linearly from pretest scores is removed from the posttest score. Compared with change score, residualized change is not a more correct measure in that it might remove some important and genuine change in the subject. The residualized score helps identify individuals who changed more (or less) than expected.

The true residual-gain score is defined as the expected value of raw residual-gain score over many observations on the same person. It is the residual obtained for a subject in the population linear regression of true final status on true initial status. The true residual gain is obtained by extending Lord's multiple-regression approach and can be further improved by including $W$ or $Z$ information.

In spite of all the developments in the estimation of change or residual change scores, researchers are often warned not to use these change scores. Cronbach and Furby summarized four diverse research issues in the measurement-of-change literature that

may appear to require change scores, including (1) providing a measure of individual change, (2) investigating correlates or predictors of change (or change rate), (3) identifying individuals with slowest change rate for further special treatment, and (4) providing an indicator of a construct that can serve as independent, dependent, or covariate variables. However, for most of these questions, the estimation of change scores is unnecessary and is at best inferior to other methods of analysis. One example is one-group studies of the pretest–treatment–posttest form. The least squares estimate of the mean true gain is simply the difference between mean observed pretest and posttest scores. Hypothesis testing and estimation related to treatment effect should be addressed directly to observed sample means. This suggests that the analysis of change does not need estimates of true change scores. It is more appropriate to directly use raw score mean vectors, covariance matrices, and estimated reliabilities. Cronbach and Furby recommended that, when investigators ask questions in which change scores appear to be the natural measure to be obtained, the researchers should frame their questions in other ways.

*Feifei Ye*

***See also*** Gain Scores, Analysis of; Longitudinal Design; Pretest–Posttest Design

**Further Readings**

Chan, D. (2003). Data analysis and modeling longitudinal processes. *Group & Organization Management, 28,* 341–365.

Cronbach, L. J., & Furby, L. (1970). How should we measure "change"—or should we? *Psychological Bulletin, 74*(1), 68–80.

Linn, R. L., & Slinde, J. A. (1977). The determination of the significance of change between pre- and posttesting periods. *Review of Educational Research, 47*(1), 121–150.

# Chi-Square Test

The chi-square test is a nonparametric test of the statistical significance of a relation between two nominal or ordinal variables. Because a chi-square analyzes grosser data than do parametric tests such as *t* tests and analyses of variance (ANOVAs), the chi-square test can report only whether groups *in a sample* are significantly different in some measured attribute or behavior; it does not allow one to generalize from the sample to the population from which it was drawn. Nonetheless, because chi-square is less "demanding" about the data it will accept, it can be used in a wide variety of research contexts. This entry focuses on the application, requirements, computation, and interpretation of the chi-square test, along with its role in determining associations among variables.

## Bivariate Tabular Analysis

Though one can apply the chi-square test to a single variable and judge whether the frequencies for each category are equal (or as expected), a chi-square is applied most commonly to frequency results reported in bivariate tables, and interpreting bivariate tables is crucial to interpreting the results of a chi-square test. Bivariate tabular analysis (sometimes called *crossbreak analysis*) is used to understand the relationship (if any) between two variables. For example, if a researcher wanted to know whether there is a relationship between the gender of U.S. undergraduates at a particular university and their footwear preferences, he or she might ask male and female students (selected as randomly as possible), "On average, do you prefer to wear sandals, sneakers, leather shoes, boots, or something else?" In this example, the independent variable is gender and the dependent variable is footwear preference. The independent variable is the quality or characteristic that the researcher hypothesizes helps to predict or explain some other characteristic or behavior (the dependent variable). Researchers control the independent variable (in this example, by sampling males and females) and elicit and measure the dependent variable to test their hypothesis that there is some relationship between the two variables.

To see whether there is a systematic relationship between gender of undergraduates at University X and reported footwear preferences, the results could be summarized in a table as shown in Table 1.

Each cell in a bivariate table represents the intersection of a value on the independent variable and a value on the dependent variable by showing

**Table 1** Male and Female Undergraduate Footwear Preferences at University X (Raw Frequencies)

| *Group* | *Sandals* | *Sneakers* | *Leather Shoes* | *Boots* | *Other* |
|---|---|---|---|---|---|
| Male | 6 | 17 | 13 | 9 | 5 |
| Female | 13 | 5 | 7 | 16 | 9 |

**Table 2** Male and Female Undergraduate Footwear Preferences at University X (Percentages)

| *Group* | *Sandals* | *Sneakers* | *Leather Shoes* | *Boots* | *Other* | *N* |
|---|---|---|---|---|---|---|
| Male | 12 | 34 | 26 | 18 | 10 | 50 |
| Female | 26 | 10 | 14 | 32 | 18 | 50 |

how many times that combination of values was observed in the sample being analyzed. Typically, in constructing bivariate tables, values on the independent variable are arrayed on the vertical axis, while values on the dependent variable are arrayed on the horizontal axis. This allows one to read "across," from values on the independent variable to values on the dependent variable. (Remember, an observed relationship between two variables is not necessarily causal.)

Reporting and interpreting bivariate tables is most easily done by converting raw frequencies (in each cell) into percentages of each cell within the categories of the independent variable. Percentages basically standardize cell frequencies as if there were 100 subjects or observations in each category of the independent variable. This is useful for comparing across values on the independent variable if the raw row totals are close to or more than 100, but increasingly dangerous as raw row totals become smaller. (When reporting percentages, one should indicate total *N* at the end of each row or independent variable category.)

Table 2 shows that in this sample roughly twice as many women as men preferred sandals and boots, about 3 times more men than women preferred sneakers, and twice as many men as women preferred leather shoes. One might also infer from the "Other" category that female students in this sample had a broader range of footwear preferences than did male students.

Converting raw observed values or frequencies into percentages allows one to see patterns in the data more easily, but how confident can one be that those apparent patterns actually reflect a systematic relationship in the sample between the variables (between gender and footwear preference, in this example) and not just a random distribution?

## Chi-Square Requirements

The chi-square test of statistical significance is a series of mathematical formulas that compare the actual observed frequencies of the two variables measured in a sample with the frequencies one would expect if there were no relationship at all between those variables. That is, chi-square assesses whether the actual results are different enough from the null hypothesis to overcome a certain probability that they are due to sampling error, randomness, or a combination.

Because chi-square is a nonparametric test, it does not require the sample data to be at an interval level of measurement and more or less normally distributed (as parametric tests such as *t* tests do), although it does rely on a weak *assumption* that each variable's values are normally distributed in the population from which the sample is drawn. But chi-square, while forgiving, does have some requirements:

1. *Chi-square is most appropriate for analyzing relationships among nominal and ordinal variables.* A nominal variable (sometimes called a categorical variable) describes an attribute in terms of mutually exclusive, nonhierarchically related categories, such as gender and footwear preference. Ordinal variables measure an attribute (such as military rank) that subjects may have more or less of but that cannot be measured in equal increments on a scale. (Results from interval variables, such as scores on a test, would have to first be grouped before they could "fit" into a bivariate table and be analyzed with chi-square; this grouping loses much of the incremental information of the original scores, so interval data are usually analyzed using parametric tests such as ANOVAs and *t* tests. The relationship between two ordinal variables is usually best analyzed with a Spearman rank order correlation.)

2. *The sample must be randomly drawn from the population.*

3. *Data must be reported in raw frequencies—not, for example, in percentages.*

4. *Measured variables must be independent of each other.* Any observation must fall into *only one* category or value on each variable, and no category can be inherently dependent on or influenced by another.

5. *Values and categories on independent and dependent variables must be mutually exclusive and exhaustive.* In the footwear data, each subject is counted only once, as either male or female and as preferring sandals, sneakers, leather shoes, boots, or other kinds of footwear. For some variables, no "other" category may be needed, but often "other" ensures that the variable has been exhaustively categorized. (Some kinds of analysis may require an "uncodable" category.) In any case, the results for the whole sample must be included.

6. *Expected (and observed) frequencies cannot be too small.* Chi-square is based on the expectation that *within any category,* sample frequencies are normally distributed about the expected population value. Since frequencies of occurrence cannot be negative, the distribution cannot be normal when expected population values are close to zero (because the sample frequencies cannot be much *below* the expected frequency while they *can* be much above it). When expected frequencies are large, there is no problem with the assumption of normal distribution, but the smaller the expected frequencies, the less valid the results of the chi-square test. In addition, because some of the mathematical formulas in chi-square use division, no cell in a table can have an observed raw frequency of zero.

The following minimums should be obeyed:

> For a $1 \times 2$ or $2 \times 2$ table, expected frequencies in each cell should be at least 5.
>
> For larger tables ($2 \times 4$ or $3 \times 3$ or larger), if all expected frequencies but one are at least 5 and if the one small cell is at least 1, chi-square is appropriate. In general, the greater the degrees of freedom (i.e., the more values or categories on the independent and dependent variables), the more lenient the minimum expected frequencies threshold.

Sometimes, when a researcher finds low expected frequencies in one or more cells, a possible solution is to combine, or collapse, two cells together. (Categories on a variable cannot be excluded from a chi-square analysis; a researcher cannot arbitrarily exclude some subset of the data from analysis.) But a decision to collapse categories should be carefully motivated, preserving the integrity of the data as it was originally collected.

## Computing the Chi-Square Value

The process by which a chi-square value is computed has four steps:

1. *Setting the* p *value.* This sets the threshold of tolerance for error. That is, what odds is the researcher willing to accept that apparent patterns in the data may be due to randomness or sampling error rather than some systematic relationship between the measured variables? The answer depends largely on the research question and the consequences of being wrong. If people's lives depend on the interpretation of the results, the researcher might want to take only one chance in 100,000 (or 1,000,000) of making an erroneous claim. But if the stakes are smaller, he or she might accept a greater risk—1 in 100 or 1 in 20. To minimize any temptation for post hoc compromise of scientific standards, researchers should explicitly motivate their threshold *before* they perform any test of statistical significance. For the footwear study, we will set a probability of error threshold of 1 in 20, or $p < .05$.

2. *Totaling all rows and columns.* See Table 3.

3. *Deriving the expected frequency of each cell.* Chi-square operates by comparing the observed frequencies in each cell in the table to the frequencies one would expect if there were no relationship at all between the two variables in the populations from which the sample is drawn (the null

**Table 3** Male and Female Undergraduate Footwear Preferences at University X: Observed Frequencies With Row and Column Totals

| | | | *Leather* | | | |
|---|---|---|---|---|---|---|
| *Group* | *Sandals* | *Sneakers* | *Shoes* | *Boots* | *Other* | *Total* |
| Male | 6 | 17 | 13 | 9 | 5 | 50 |
| Female | 13 | 5 | 7 | 16 | 9 | 50 |
| Total | 19 | 22 | 20 | 25 | 14 | 100 |

**Table 4** Male and Female Undergraduate Footwear Preferences at University X: Observed and Expected Frequencies

| | | | Leather | | | |
|---|---|---|---|---|---|---|
| *Group* | *Sandals* | *Sneakers* | *Shoes* | *Boots* | *Other* | *Total* |
| Male: | 6 | 17 | 13 | 9 | 5 | 50 |
| observed | 9.5 | 11 | 10 | 12.5 | 7 | |
| expected | | | | | | |
| Female: | 13 | 5 | 7 | 16 | 9 | 50 |
| observed | 9.5 | 11 | 10 | 12.5 | 7 | |
| expected | | | | | | |
| Total | 19 | 22 | 20 | 25 | 14 | 100 |

hypothesis). The null hypothesis—the "all other things being equal" scenario—is derived from the observed frequencies as follows: The expected frequency in each cell is the product of that cell's row total multiplied by that cell's column total, divided by the sum total of all observations. So, to derive the expected frequency of the "Males who prefer Sandals" cell, multiply the top row total (50) by the first column total (19) and divide that product by the sum total (100): $((50 \times 19)/100) = 9.5$. The logic of this is that we are deriving the expected frequency of each cell from the union of the total frequencies of the relevant values on each variable (in this case, Male and Sandals), as a proportion of all observed frequencies in the sample. This calculation produces the expected frequency of each cell, as shown in Table 4.

Now a comparison of the observed results with the results one would expect if the null hypothesis were true is possible. (Because the sample includes the same number of male and female subjects, the male and female expected scores are the same. This will not be the case with unbalanced samples.) This table can be informally analyzed, comparing observed and expected frequencies in each cell (e.g., males prefer sandals less than expected), across values on the independent variable (e.g., males prefer sneakers more than expected, females less than expected), or across values on the dependent variable (e.g., females prefer sandals and boots more than expected, but sneakers and shoes less than expected). But some way to measure how different the observed results are from the null hypothesis is needed.

4. *Measuring the size of the difference between the observed and expected frequencies in each cell.* To do this, calculate the difference between the observed and expected frequency in each cell, square that difference, and then divide that product by the difference itself. The formula can be expressed as $(O - E)^2/E$.

Squaring the difference ensures a positive number, so that one ends up with an absolute value of differences. (If one did not work with absolute values, the positive and negative differences across the entire table would always add up to 0.) Dividing the squared difference by the expected frequency essentially removes the expected frequency from the equation, so that the remaining measures of observed versus expected difference are comparable across all cells.

So, for example, the difference between observed and expected frequencies for the Male–Sandals preference is calculated as follows:

1. Observed (6) − Expected (9.5) = Difference (−3.5)
2. Difference ( − 3.5) squared = 12.25
3. Difference squared (12.25)/Expected (9.5) = 1.289

The sum of all products of this calculation on each cell is the total chi-square value for the table: 14.026.

## Interpreting the Chi-Square Value

Now the researcher needs some criterion against which to measure the table's chi-square value in order to tell whether it is significant (relative to the *p* value that has been motivated). The researcher needs to know the probability of getting a chi-square value of a given minimum size even if the variables are not related at all in the sample. That is, the researcher needs to know how much larger than zero (the chi-square value of the null hypothesis) the table's chi-square value must be before the null hypothesis can be rejected with confidence. The probability depends in part on the complexity and sensitivity of the variables, which are reflected in the degrees of freedom of the table from which the chi-square value is derived.

A table's degrees of freedom (*df*) can be expressed by this formula: $df = (r - 1)(c - 1)$.

That is, a table's degrees of freedom equals the number of rows in the table minus 1, multiplied by the number of columns in the table minus 1. (For $1 \times 2$ tables, $df = k - 1$, where $k =$ number of values or categories on the variable.) Different chi-square thresholds are set for relatively gross comparisons ($1 \times 2$ or $2 \times 2$) versus finer comparisons. (For Table 1, $df = (2 - 1)(5 - 1) = 4$.)

In a statistics book, the sampling distribution of chi-square (also known as *critical values of chi-square*) is typically listed in an appendix. The researcher can read down the column representing his or her previously chosen probability of error threshold (e.g., $p < .05$) and across the row representing the degrees of freedom in his or her table. If the researcher's chi-square value is larger than the critical value in that cell, his or her data represent a statistically significant relationship between the variables in the table. (In statistics software programs, all the computations are done for the researcher, and he or she is given the exact $p$ value of the results.)

Table 1's chi-square value of 14.026, with $df = 4$, is greater than the related critical value of 9.49 (at $p = .05$), so the null hypothesis can be rejected, and the claim that the male and female undergraduates at University X in this sample differ in their (self-reported) footwear preferences and that there is some relationship between gender and footwear preferences (in this sample) can be affirmed.

A statistically significant chi-square value indicates the degree of confidence a researcher may hold that the relationship between variables in the sample is systematic and not attributable to random error. It does not help the researcher to interpret the nature or explanation of that relationship; that must be done by other means (including bivariate tabular analysis and qualitative analysis of the data).

## Measures of Association

By itself, statistical significance does not ensure that the relationship is theoretically or practically important or even very large. A large enough sample may demonstrate a statistically significant relationship between two variables, but that relationship may be a trivially weak one. The relative strength of association of a statistically significant relationship between the variables in the data can be derived from a table's chi-square value.

For tables larger than $2 \times 2$ (such as Table 1), a measure called Cramer's V is derived by the following formula (where $N =$ the total number of observations, and $k =$ the smaller of the number of rows or columns):

Cramer's V = the square root of
(chi-square divided by ($N$ times ($k$ minus 1)))

So, for ($2 \times 5$) Table 1, Cramer's V can be computed as follows:

1. $N(k - 1) = 100\ (2 - 1) = 100$
2. Chi-square/100 = 14.026/100 = 0.14
3. Square root of 0.14 = 0.37

The product is interpreted as a Pearson correlation coefficient ($r$). (For $2 \times 2$ tables, a measure called *phi* is derived by dividing the table's chi-square value by $N$ (the total number of observations) and then taking the square root of the product. Phi is also interpreted as a Pearson $r$.) Also, $r^2$ is a measure called *shared variance*. Shared variance is the portion of the total representation of the variables measured in the sample data that is accounted for by the relationship measured by the chi-square. For Table 1, $r^2 = .137$, so approximately 14% of the total footwear preference story is accounted for by gender.

A measure of association like Cramer's V is an important benchmark of just "how much" of the phenomenon under investigation has been accounted for. For example, Table 1's Cramer's V of 0.37 ($r^2 = .137$) means that there are one or more variables remaining that, cumulatively, account for at least 86% of footwear preferences. This measure, of course, does not begin to address the nature of the relation(s) among these variables, which is a crucial part of any adequate explanation or theory.

*Jeff Connor-Linton*

***See also*** Correlation; Critical Value; Degrees of Freedom; Dependent Variable; Expected Value; Frequency Table; Hypothesis; Independent Variable; Nominal Scale; Nonparametric Statistics; Ordinal Scale; $p$ Value; $R^2$; Random Error; Random Sampling; Sampling Error; Variable; Yates's Correction.

**Further Readings**

Berman, E. M. (2006). *Essential statistics for public managers and policy analysts* (2nd ed.). Washington, DC: CQ Press.
Field, A. (2005). *Discovering statistics using SPSS* (2nd ed.). London: Sage.
Hatch, E., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics*. Rowley, MA: Newbury House.
Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.
Sokal, R. R., & Rohlf, F. J. (1995). *Biometry: The principles and practice of statistics in biological research* (3rd ed.). New York: W. H. Freeman.
Welkowitz, J., Ewen, R. B., & Cohen, J. (1982). *Introductory statistics for the behavioral sciences* (3rd ed.). New York: Academic Press.

# Classical Test Theory

Measurement is the area of quantitative social science that is concerned with ascribing numbers to individuals in a meaningful way. Measurement is distinct from statistics, though measurement theories are grounded in applications of statistics. Within measurement there are several theories that allow us to talk about the quality of measurements taken. Classical test theory (CTT) can arguably be described as the first formalized theory of measurement and is still the most commonly used method of describing the characteristics of assessments. With test theories, the term *test* or *assessment* is applied widely. Surveys, achievement tests, intelligence tests, psychological assessments, writing samples graded with rubrics, and innumerable other situations in which numbers are assigned to individuals can all be considered tests. The terms *test, assessment, instrument,* and *measure* are used interchangeably in this discussion of CTT. After a brief discussion of the early history of CTT, this entry provides a formal definition and discusses CTT's role in reliability and validity.

## Early History

Most of the central concepts and techniques associated with CTT (though it was not called that at the time) were presented in papers by Charles Spearman in the early 1900s. One of the first texts to codify the emerging discipline of measurement and CTT was Harold Gulliksen's *Theory of Mental Tests*. Much of what Gulliksen presented in that text is used unchanged today. Other theories of measurement (generalizability theory, item response theory) have emerged that address some known weaknesses of CTT (e.g., homoscedasticity of error along the test score distribution). However, the comparative simplicity of CTT and its continued utility in the development and description of assessments have resulted in CTT's continued use. Even when other test theories are used, CTT often remains an essential part of the development process.

## Formal Definition

CTT relies on a small set of assumptions. The implications of these assumptions build into the useful CTT paradigm. The fundamental assumption of CTT is found in the equation

$$X = T + E, \tag{1}$$

where $X$ represents an observed score, $T$ represents true score, and $E$ represents error of measurement.

The concept of the true score, $T$, is often misunderstood. A true score, as defined in CTT, does not have any direct connection to the construct that the test is intended to measure. Instead, the true score represents the number that is the *expected* value for an individual based on this specific test. Imagine that a test taker took a 100-item test on world history. This test taker would get a score, perhaps an 85. If the test was a multiple-choice test, probably some of those 85 points were obtained through guessing. If given the test again, the test taker might guess better (perhaps obtaining an 87) or worse (perhaps obtaining an 83). The causes of differences in observed scores are not limited to guessing. They can include anything that might affect performance: a test taker's state of being (e.g., being sick), a distraction in the testing environment (e.g., a humming air conditioner), or careless mistakes (e.g., misreading an essay prompt).

Note that the true score is theoretical. It can never be observed directly. Formally, the true score is assumed to be the expected value (i.e., average)

of $X$, the observed score, over an infinite number of independent administrations. Even if an examinee could be given the same test numerous times, the administrations would not be independent. There would be practice effects, or the test taker might learn more between administrations.

Once the true score is understood to be an average of observed scores, the rest of Equation 1 is straightforward. A scored performance, $X$, can be thought of as deviating from the true score, $T$, by some amount of error, $E$. There are additional assumptions related to CTT: $T$ and $E$ are uncorrelated ($\rho_{TE} = 0$), error on one test is uncorrelated with the error on another test ($\rho_{E_1E_2} = 0$), and error on one test is uncorrelated with the true score on another test ($\rho_{E_1E_2} = 0$). These assumptions will not be elaborated on here, but they are used in the derivation of concepts to be discussed.

## Reliability

If $E$ tends to be large in relation to $T$, then a test result is inconsistent. If $E$ tends to be small in relation to $T$, then a test result is consistent. The major contribution of CTT is the formalization of this concept of consistency of test scores. In CTT, test score consistency is called *reliability*. Reliability provides a framework for thinking about and quantifying the consistency of a test. Even though reliability does not directly address constructs, it is still fundamental to measurement. The often used bathroom scale example makes the point. Does a bathroom scale accurately reflect weight? Before establishing the scale's accuracy, its consistency can be checked. If one were to step on the scale and it said 190 pounds, then step on it again and it said 160 pounds, the scale's lack of utility could be determined based strictly on its inconsistency.

Now think about a formal assessment. Imagine an adolescent was given a graduation exit exam and she failed, but she was given the same exam again a day later and she passed. She did not learn in one night everything she needed in order to graduate. The consequences would have been severe if she had been given only the first opportunity (when she failed). If this inconsistency occurred for most test takers, the assessment results would have been shown to have insufficient consistency (i.e., reliability) to be useful. Reliability coefficients allow for the quantification of consistency so that decisions about utility can be made.

$T$ and $E$ cannot be observed directly for individuals, but the relative contribution of these components to $X$ is what defines reliability. Instead of direct observations of these quantities, group-level estimates of their variability are used to arrive at an estimate of reliability. Reliability is usually defined as

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_T^2} + \sigma_{E^2}, \qquad (2)$$

where $X$, $T$, and $E$ are defined as before. From this equation one can see that reliability is the ratio of true score variance to observed score variance within a sample. Observed score variance is the only part of this equation that is observable. There are essentially three ways to estimate the unobserved portion of this equation. These three approaches correspond to three distinct types of reliability: *stability reliability, alternate-form reliability,* and *internal consistency.*

All three types of reliability are based on correlations that exploit the notion of parallel tests. Parallel tests are defined as tests having equal true scores ($T = T'$) and equal error variance ($\sigma_E^2 = \sigma_{E'}^2$) and meeting all other assumptions of CTT. Based on these assumptions, it can be derived (though it will not be shown here) that the correlation between parallel tests provides an estimate of reliability (i.e., the ratio of true score variance to observed score variance). The correlation among these two scores is the observed reliability coefficient ($r_{XX'}$). The correlation can be interpreted as the proportion of variance in observed scores that is attributable to true scores; thus $r_{XX'} = r_{XT}^2$. The type of reliability (stability reliability, alternate-form reliability, internal consistency) being estimated is determined by how the notion of a parallel test is established.

The most straightforward type of reliability is stability reliability. Stability reliability (often called test–retest reliability) is estimated by having a representative sample of the intended testing population take the same instrument twice. Because the same test is being used at both measurement opportunities, the notion of parallel tests is easy to support. The same test is used on both occasions, so true scores and error variance should be the same. Ideally, the two measurement opportunities

will be close enough that examinees have not changed (learned or developed on the relevant construct). You would not, for instance, want to base your estimates of stability reliability on pretest–intervention–posttest data. The tests should also not be given too close together in time; otherwise, practice or fatigue effects might influence results. The appropriate period to wait between testing will depend on the construct and purpose of the test and may be anywhere from minutes to weeks.

Alternate-form reliability requires that each member of a representative sample respond on two alternate assessments. These alternate forms should have been built to be purposefully parallel in content and scores produced. The tests should be administered as close together as is practical, while avoiding fatigue effects. The correlation among these forms represents the alternate-form reliability. Higher correlations provide more confidence that the tests can be used interchangeably (comparisons of means and standard deviations will also influence this decision).

Having two administrations of assessments is often impractical. Internal consistency reliability methods require only a single administration. There are two major approaches to estimating internal consistency: split half and coefficient alpha. Split-half estimates are easily understood but have mostly been supplanted by the use of coefficient alpha. For a split-half approach, a test is split into two halves. This split is often created by convenience (e.g., all odd items in one half, all even items in the other half). The split can also be made more methodically (e.g., balancing test content and item types). Once the splits are obtained, the scores from the two halves are correlated. Conceptually, a single test is being used to create an estimate of alternate-form reliability. For reasons not discussed in this entry, the correlation from a split-half method will underestimate reliability unless it is corrected.

The *Spearman–Brown prophecy formula* for predicting the correlation that would have been obtained if each half had been as long as the full-length test is given by

$$\rho_{XX'} = \frac{2\rho_{AB}}{1 + \rho_{AB}}, \tag{3}$$

where $\rho_{AB}$ is the original correlation between the test halves and $\rho_{XX'}$ is the corrected reliability. If, for example, two test halves were found to have a correlation of .60, the actual reliability would be

$$\rho_{XX'} = \frac{2(.60)}{1 + .60} = \frac{1.20}{1.60} = .75.$$

As with all the estimates of reliability, the extent to which the two measures violate the assumption of parallel tests determines the accuracy of the result.

One of the drawbacks of using the split-half approach is that it does not produce a unique result. Other splits of the test will produce (often strikingly) different estimates of the internal consistency of the test. This problem is ameliorated by the use of single internal consistency coefficients that provide information similar to split-half reliability estimates. Coefficient alpha (often called Cronbach's alpha) is the most general (and most commonly used) estimate of internal consistency. The Kuder–Richardson 20 (KR-20) and 21 (KR-21) are reported sometimes, but these coefficients are special cases of coefficient alpha and do not require separate discussion. The formula for coefficient alpha is

$$\rho_{XX'} \geq \alpha = \frac{k}{k-1}\left(1 - \frac{\Sigma\sigma_i^2}{\sigma_X^2}\right), \tag{4}$$

where $k$ is the number of items, $\sigma_i^2$ is the variance of item $i$, and $\sigma_X^2$ is the variance of test $X$. Coefficient alpha is equal to the average of all possible split-half methods computed using Phillip Rulon's method (which uses information on the variances of total scores and the differences between the split-half scores). Coefficient alpha is considered a conservative estimate of reliability that can be interpreted as a lower bound to reliability. Alpha is one of the most commonly reported measures of reliability because it requires only a single test administration.

Although all three types of reliability are based on the notions of parallel tests, they are influenced by distinct types of error and, therefore, are not interchangeable. Stability reliability measures the extent to which occasions influence results. Errors associated with this type of reliability address how small changes in examinees or the testing environment impact results. Alternate-form reliability addresses how small differences in different versions of tests may impact results. Internal consistency addresses the way

in which heterogeneity of items limits the information provided by an assessment. In some sense, any one coefficient may be an overestimate of reliability (even alpha, which is considered to be conservative). Data collected on each type of reliability would yield three different coefficients that may be quite different from one another. Researchers should be aware that a reliability coefficient that included all three sources of error at once would likely be lower than a coefficient based on any one source of error. *Generalizability theory* is an extension of CTT that provides a framework for considering multiple sources of error at once. Generalizability coefficients will tend to be lower than CTT-based reliability coefficients but may more accurately reflect the amount of error in measurements. Barring the use of generalizability theory, a practitioner must decide what types of reliability are relevant to his or her research and make sure that there is evidence of that type of consistency (i.e., through consultation of appropriate reliability coefficients) in the test results.

## Validity Investigations and Research With Classical Test Theory

CTT is mostly a framework for investigating reliability. Most treatments of CTT also include extensive descriptions of validity; however, similar techniques are used in the investigation of validity whether or not CTT is the test theory being employed. In short hand, validity addresses the question of whether test results provide the intended information. As such, validity evidence is primary to any claim of utility. A measure can be perfectly reliable, but it is useless if the intended construct is not being measured. Returning to the bathroom scale example, if a scale always describes one adult of average build as 25 pounds and second adult of average build as 32 pounds, the scale is reliable. However, the scale is not accurately reflecting the construct *weight*.

Many sources provide guidance about the importance of validity and frameworks for the types of data that constitute validity evidence. More complete treatments can be found in Samuel Messick's many writings on the topic. A full treatment is beyond the scope of this discussion. Readers unfamiliar with the current unified understanding of validity should consult a more complete reference; the topics to be addressed here might convey an overly simplified (and positivist) notion of validity.

Construct validity is the overarching principle in validity. It asks, Is the correct construct being measured? One of the principal ways that construct validity is established is by a demonstration that tests are associated with criteria or other tests that purport to measure the same (or related) constructs. The presence of strong associations (e.g., correlations) provides evidence of construct validity. Additionally, research agendas may be established that investigate whether test performance is affected by things (e.g., interventions) that should influence the underlying construct. Clearly, the collection of construct validity evidence is related to general research agendas. When one conducts basic or applied research using a quantitative instrument, two things are generally confounded in the results: (1) the validity of the instrument for the purpose being employed in the research and (2) the correctness of the research hypothesis. If a researcher fails to support his or her research hypothesis, there is often difficulty determining whether the result is due to insufficient validity of the assessment results or flaws in the research hypothesis.

One of CTT's largest contributions to our understanding of both validity investigations and research in general is the notion of reliability as an upper bound to a test's association with another measure. From Equation 1, observed score variance is understood to comprise a true score, $T$, and an error term, $E$. The error term is understood to be random error. Being that it is random, it cannot correlate with anything else. Therefore, the true score component is the only component that may have a nonzero correlation with another variable. The notion of reliability as the ratio of true score variance to observed score variance (Equation 2) makes the idea of reliability as an upper bound explicit. Reliability is the proportion of systematic variance in observed scores. So even if there were a perfect correlation between a test's true score and a perfectly measured criterion, the observed correlation could be no larger than the square root of the reliability of the test. If both quantities involved in the correlation have less than perfect

reliability, the observed correlation would be even smaller. This logic is often exploited in the reverse direction to provide a correction for attenuation when there is a lack of reliability for measures used. The correction for attenuation provides an upper bound estimate for the correlation among true scores. The equation is usually presented as

$$\rho_{T_X T_Y} = \frac{\rho_{XY}}{\sqrt{\rho_{XX'}}\sqrt{\rho_{YY'}}} \quad (5)$$

where $\rho_{XY}$ is the observed correlation between two measures and $\rho_{XX'}$ and $\rho_{YY'}$ are the respective reliabilities of the two measures. This correction for attenuation provides an estimate for the correlation among the underlying constructs and is therefore useful whenever establishing the magnitude of relationships among constructs is important. As such, it applies equally well to many research scenarios as it does to attempts to establish construct validity. For instance, it might impact interpretation if a researcher realized that an observed correlation of .25 between two instruments which have internal consistencies of .60 and .65 represented a potential true score correlation of

$$\rho_{T_X T_Y} = \frac{.25}{\sqrt{.60}\sqrt{.65}} = .40$$

The corrected value should be considered an upper bound estimate for the correlation if all measurement error were purged from the instruments. Of course, such perfectly reliable measurement is usually impossible in the social sciences, so the corrected value will generally be considered theoretical.

### Group-Dependent Estimates

Estimates in CTT are highly group dependent. Any evidence of reliability and validity is useful only if that evidence was collected on a group similar to the current target group. A test that is quite reliable and valid with one population may not be with another. To make claims about the utility of instruments, testing materials (e.g., test manuals or technical reports) must demonstrate the appropriateness and representativeness of the samples used.

### Final Thoughts

CTT provides a useful framework for evaluating the utility of instruments. Researchers using quantitative instruments need to be able to apply fundamental concepts of CTT to evaluate the tools they use in their research. Basic reliability analyses can be conducted with the use of most commercial statistical software (e.g., SPSS—an IBM company, formerly called PASW® Statistics—and SAS).

*John T. Willse*

***See also*** Reliability; "Validity"

#### Further Readings

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Harcourt Brace Jovanovich.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15,* 72–101.

Traub, R. E. (1987). Classical test theory in historical perspective. *Educational Measurement: Issues & Practice, 16*(4), 8–14.

## Clinical Significance

In treatment outcome research, statistically significant changes in symptom severity or end-state functioning have traditionally been used to demonstrate treatment efficacy. In more recent studies, the *effect size,* or magnitude of change associated with the experimental intervention, has also been an important consideration in data analysis and interpretation. To truly understand the impact of a research intervention, it is essential for the investigator to adjust the lens, or

"zoom out," to also examine other signifiers of change. Clinical significance is one such marker and refers to the meaningfulness, or impact of an intervention, on clients and others in their social environment. An intervention that is clinically significant must demonstrate substantial or, at least, reasonable benefit to the client or others, such as family members, friends, or coworkers. Benefits gained, actual or perceived, must be weighed against the costs of the intervention. These costs may include financial, time, or family burden. Some researchers use the term *practical significance* as a synonym for clinical significance because both terms consider the import of a research finding in everyday life. However, there are differences in the usage of the two terms. *Clinical significance* is typically constrained to treatment outcome or prevention studies whereas *practical significance* is used broadly across many types of psychological research, including cognitive neuroscience, developmental psychology, environmental psychology, and social psychology. This entry discusses the difference between statistical and clinical significance and describes methods for measuring clinical significance.

## Statistical Significance Versus Clinical Significance

Statistically significant findings do not always correspond to the client's phenomenological experience or overall evaluation of beneficial impact. First, a research study may have a small effect size yet reveal statistically significant findings due to high power. This typically occurs when a large sample size has been used. Nevertheless, the clinical significance of the research findings may be trivial from the research participants' perspective. Second, in other situations, a moderate effect size may yield statistically significant results, yet the pragmatic benefit to the client in his or her everyday life is questionable. For example, children diagnosed with attention-deficit/hyperactivity disorder may participate in an intervention study designed to increase concentration and reduce disruptive behavior. The investigators conclude that the active intervention was beneficial on the basis of significant improvement on Connors Ratings Scale scores as well as significantly improved performance on a computerized test measuring sustained attention. The data interpretation is correct from a statistical point of view. However, the majority of parents view the intervention as inconsequential because their children continue to evidence a behavioral problem that disrupts home life. Moreover, most of the treated children see the treatment as "a waste of time" because they are still being teased or ostracized by peers. Despite significant sustained attention performance improvements, classroom teachers also rate the experimental intervention as ineffective because they did not observe meaningful changes in academic performance.

A third scenario is the case of null research findings in which there is also an inconsistency between statistical interpretation and the client or family perspective. For instance, an experimental treatment outcome study is conducted with adult trauma survivors compared with treatment as usual. Overall, the treatment-as-usual group performed superiorly to the new intervention. In fact, on the majority of posttraumatic stress disorder (PTSD) measures, the experimental group evidenced no statistical change from pre- to posttest. Given the additional costs of the experimental intervention, the investigators may decide that it is not worth further investigation. However, qualitative interviews are conducted with the participants. The investigators are surprised to learn that most participants receiving the intervention are highly satisfied although they continue to meet PTSD diagnostic criteria. The interviews demonstrate clinical significance among the participants, who perceive a noticeable reduction in the intensity of daily dissociative symptoms. These participants see the experimental intervention as quite beneficial in terms of facilitating tasks of daily living and improving their quality of life.

When planning a research study, the investigator should consider *who* will evaluate clinical significance (client, family member, investigator, original treating therapist) and *what* factor(s) are important to the evaluator (changes in symptom severity, functioning, personal distress, emotional regulation, coping strategies, social support resources, quality of life, etc.). An investigator should also consider the cultural context and the

rater's cultural expertise in the area under examination. Otherwise, it may be challenging to account for unexpected results. For instance, a 12-week mindfulness-based cognitive intervention for depressed adults is initially interpreted as successful in improving mindfulness skills, on the basis of statistically significant improvements on two dependent measures: scores on a well-validated self-report mindfulness measure and attention focus ratings by the Western-trained research therapists. The investigators are puzzled that most participants do not perceive mindfulness skill training as beneficial; that is, the training has not translated into noticeable improvements in depression or daily life functioning. The investigators request a second evaluation by two mindfulness experts—a non-Western meditation practitioner versed in traditional mindfulness practices and a highly experienced yoga practitioner. The mindfulness experts independently evaluate the research participants and conclude that culturally relevant mindfulness performance markers (postural alignment, breath control, and attentional focus) are very weak among participants who received the mindfulness intervention.

### Measures

Clinical significance may be measured several ways, including subjective evaluation, absolute change (did the client evidence a complete return to premorbid functioning or how much has an individual client changed across the course of treatment without comparison with a reference group), comparison method, or societal impact indices. In most studies, the comparison method is the most typically employed strategy for measuring clinical significance. It may be used for examining whether the client group returns to a normative level of symptoms or functioning at the conclusion of treatment. Alternatively, the comparison method may be used to determine whether the experimental group statistically differs from an impaired group at the conclusion of treatment even if the experimental group has not returned to premorbid functioning. For instance, a treatment outcome study is conducted with adolescents diagnosed with bulimia nervosa. After completing the intervention, the investigators may consider whether the level of body image dissatisfaction is comparable to that of a normal comparison group with no history of eating disorders, or the study may examine whether binge–purge end-state functioning statistically differs from a group of individuals currently meeting diagnostic criteria for bulimia nervosa.

The *Reliable Change Index,* developed by Neil Jacobson and colleagues, and *equivalence testing* are the most commonly used comparison method strategies. These comparative approaches have several limitations, however, and the reader is directed to the Further Readings for more information on the conceptual and methodological issues.

The tension between researchers and clinical practitioners will undoubtedly lessen as clinical significance is foregrounded in future treatment outcome studies. Quantitative and qualitative measurement of clinical significance will be invaluable in deepening our understanding of factors and processes that contribute to client transformation.

*Carolyn Brodbeck*

***See also*** Effect Size, Measures of; Power; Significance, Statistical

#### Further Readings

Beutler, L. E., & Moleiro, C. (2001). Clinical versus reliable and significant change. *Clinical Psychology: Science and Practice, 8,* 441–445.

Jacobson, N. S., Follette, W. C., & Revenstrof, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior Therapy, 15,* 335–352.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting 19.*

Jensen, P. S. (2001). Clinical equivalence: A step, misstep, or just a misnomer? *Clinical Psychology: Science & Practice, 8,* 436–440.

Kendall, P. C., Marrs-Garcia, A., Nath, S. R., & Sheldrick, R. C. (1999). Normative comparisons for the evaluation of clinical significance. *Journal of Consulting & Clinical Psychology, 67,* 285–299.

## Clinical Trial

A clinical trial is a prospective study that involves human subjects in which an intervention is to be

evaluated. In a clinical trial, subjects are followed from a well-defined starting point or baseline. The goal of a clinical trial is to determine whether a cause-and-effect relationship exists between the intervention and response. Examples of interventions used in clinical trials include drugs, surgery, medical devices, and education and subject management strategies. In each of these cases, clinical trials are conducted to evaluate both the beneficial and harmful effects of the new intervention on human subjects before it is made available to the population of interest. Special considerations for conducting clinical trials include subject safety and informed consent, subject compliance, and intervention strategies to avoid bias. This entry describes the different types of clinical trials and discusses ethics in relation to clinical trials.

## Drug Development Trials

Clinical trials in drug development follow from laboratory experiments, usually involving in vitro experiments or animal studies. The traditional goal of a preclinical study is to obtain preliminary information on pharmacology and toxicology. Before a new drug may be used in human subjects, several regulatory bodies, such as the internal review board (IRB), Food and Drug Administration (FDA), and data safety monitoring board, must formally approve the study. Clinical trials in drug development are conducted in a sequential fashion and categorized as Phase I, Phase II, Phase III, and Phase IV trial designs. The details of each phase of a clinical trial investigation are well defined within a document termed a *clinical trial protocol*. The FDA provides recommendations for the structure of Phase I through III trials in several disease areas.

*Phase I trials* consist primarily of healthy volunteer and participant studies. The primary objective of a Phase I trial is to determine the maximum tolerated dose. Other objectives include determining drug metabolism and bioavailability (how much drug reaches the circulation system). Phase I studies generally are short-term studies that involve monitoring toxicities in small cohorts of participants treated at consistently higher dose levels of the new drug in order to estimate the maximum tolerated dose.

*Phase II trials* build on the Phase I results in terms of which dose level or levels warrant further investigation. Phase II trials are usually fairly small-scale trials. In cancer studies, Phase II trials traditionally involve a single dose with a surrogate end point for mortality, such as change in tumor volume. The primary comparison of interest is the effect of the new regimen versus established response rates. In other diseases, such as cardiology, Phase II trials may involve multiple dose levels and randomization. The primary goals of a Phase II trial are to determine the optimal method of administration and examine the potential efficacy of a new regimen. Phase II trials generally have longer follow-up times than do Phase I trials. Within Phase II trials, participants are closely monitored for safety. In addition, pharmacokinetic, pharmacodynamic, or pharmacogenomic studies or a combination are often incorporated as part of the Phase II trial design. In many settings, two or more Phase II trials are undertaken prior to a Phase III trial.

*Phase III trials* are undertaken if the Phase II trial or trials demonstrate that the drug may be reasonably safe and potentially effective. The primary goal of a Phase III trial is to compare the effectiveness of the new treatment with that of either a placebo condition or standard of care. Phase III trials may involve long-term follow-up and many participants. The sample size is determined using precise statistical methods based on the end point of interest, the clinically relevant difference between treatment arms, and control of Type I and Type II error rates. The FDA may require two Phase III trials for approval of new drugs. The process from drug synthesis to the completion of a Phase III trial may take several years. Many investigations may never reach the Phase II or Phase III trial stage.

The gold-standard design in Phase III trials employs randomization and is double blind. That is, participants who are enrolled in the study agree to have their treatment selected by a random or pseudorandom process, and neither the evaluator nor the participants have knowledge of the true treatment assignment. Appropriate randomization procedures ensure that the assigned treatment is independent of any known or unknown prognostic

characteristic. Randomization provides two important advantages. First, the treatment groups tend to be comparable with regard to prognostic factors. Second, it provides the theoretical underpinning for a valid statistical test of the null hypothesis. Randomization and blinding are mechanisms used to minimize potential bias that may arise from multiple sources. Oftentimes blinding is impossible or infeasible due to either safety issues or more obvious reasons such as the clear side effects of the active regimen compared with a placebo control group. Randomization cannot eliminate statistical bias due to issues such as differential dropout rates, differential noncompliance rates, or any other systematic differences imposed during the conduct of a trial.

*Phase IV* clinical trials are oftentimes referred to as postmarketing studies and are generally surveillance studies conducted to evaluate long-term safety and efficacy within the new patient population.

## Nonrandomized Clinical Trials

In nonrandomized observational studies, there is no attempt to intervene in the participants' selection between the new and the standard treatments. When the results of participants who received the new treatment are compared with the results of a group of previously treated individuals, the study is considered historically controlled. Historically controlled trials are susceptible to several sources of bias, including those due to shifts in medical practice over time, changes in response definitions, and differences in follow-up procedures or recording methods. Some of these biases can be controlled by implementing a parallel treatment design. In this case, the new and the standard treatment are applied to each group of participants concurrently. This approach provides an opportunity to standardize the participant care and evaluation procedures used in each treatment group. The principal drawback of observational trials (historically controlled or parallel treatment) is that the apparent differences in outcome between the treatment groups may actually be due to imbalances in known and unknown prognostic subject characteristics. These differences in outcome can be erroneously attributed to the treatment, and thus these types of trials cannot establish a cause-and-effect relationship.

## Screening Trials

Screening trials are conducted to determine whether preclinical signs of a disease should be monitored to increase the chance of detecting individuals with curable forms of the disease. In these trials there are actually two interventions. The first intervention is the program of monitoring normal individuals for curable forms of disease. The second intervention is the potentially curative procedure applied to the diseased, detected individuals. Both these components need to be effective for a successful screening trial.

## Prevention Trials

Prevention trials evaluate interventions intended to reduce the risk of developing a disease. The intervention may consist of a vaccine, high-risk behavior modification, or dietary supplement.

## Ethics in Clinical Trials

The U.S. Research Act of 1974 established the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. The purpose of this commission was to establish ethical guidelines for clinical research. The Act also mandated that all clinical trials funded entirely or in part by the federal government be reviewed by IRBs. The Commission outlined in detail how IRBs should function and provided ethical guidelines for conducting research involving human participants. The Commission's Belmont Report articulated three important ethical principles for clinical research: respect for persons, beneficence, and justice.

The principle of respect for persons acknowledges that participants involved in clinical trials must be properly informed and always permitted to exercise their right to self-determination. The informed consent process arises from this principle. The principle of beneficence acknowledges that there is often a potential for benefit as well as harm for individuals involved in clinical trials. This principle requires that there be a favorable balance toward benefit, and any potential for harm must be minimized and justified. The principle of justice requires that the burdens and

benefits of clinical research be distributed fairly. An injustice occurs when one group in society bears a disproportionate burden of research while another reaps a disproportionate share of its benefit.

*Alan Hutson and Mark Brady*

***See also*** Adaptive Designs in Clinical Trials; Ethics in the Research Process; Group-Sequential Designs in Clinical Trials

**Further Readings**

Friedman, L. M., Furberg, C. D., & Demets, D. L. (1998). *Fundamentals of clinical trials* (3rd ed.). New York: Springer Science.
Machin, D., Day, S., & Green, S. (2006). *Textbook of clinical trials* (2nd ed.). Hoboken, NJ: Wiley.
Piantadosi, S. (2005). *Clinical trials: A methodologic perspective* (2nd ed.). Hoboken, NJ: Wiley.
Senn, S. (2008). *Statistical issues in drug development* (2nd ed.). Hoboken, NJ: Wiley.

## Cluster Sampling

A variety of sampling strategies are available in cases when setting or context create restrictions. For example, *stratified sampling* is used when the population's characteristics such as ethnicity or gender are related to the outcome or dependent variables being studied. *Simple random sampling*, in contrast, is used when there is no regard for strata or defining characteristics of the population from which the sample is drawn. The assumption is that the differences in these characteristics are normally distributed across all potential participants.

*Cluster sampling* is the selection of units of natural groupings rather than individuals. For example, in marketing research, the question at hand might be how adolescents react to a particular brand of chewing gum. The researcher may access such a population through traditional channels such as the high school but may also visit places where these potential participants tend to spend time together, such as shopping malls and movie theaters. Rather than counting each one of the adolescents' responses to a survey as one data point, the researcher would count the entire group's average as the data point. The assumption is that the group data point is a small representative of the population of all adolescents. The fact that the collection of individuals in the unit serves as the data point, rather than each individual serving as a data point, differentiates this sampling technique from most others.

An advantage of cluster sampling is that it is a great time saver and relatively efficient in that travel time and other expenses are saved. The primary disadvantage is that one can lose the heterogeneity that exists within groups by taking all in the group as a single unit; in other words, the strategy may introduce sampling error. Cluster sampling is also known as *geographical sampling* because areas such as neighborhoods become the unit of analysis.

An example of cluster sampling can be seen in a study by Michael Burton from the University of California and his colleagues, who used both stratified and cluster sampling to draw a sample from the United States Census Archives for California in 1880. These researchers emphasized that with little effort and expended resources, they obtained very useful knowledge about California in 1880 pertaining to marriage patterns, migration patterns, occupational status, and categories of ethnicity. Another example is a study by Lawrence T. Lam and L. Yang of duration of sleep and attention deficit/hyperactivity disorder among adolescents in China. The researchers used a variant of simple cluster sampling, two-stage random cluster sampling design, to assess duration of sleep.

While cluster sampling may not be the first choice given all options, it can be a highly targeted sampling method when resources are limited and sampling error is not a significant concern.

*Neil J. Salkind*

***See also*** Convenience Sampling; Experience Sampling Method; Nonprobability Sampling; Probability Sampling; Proportional Sampling; Quota Sampling; Random Sampling; Sampling Error; Stratified Sampling; Systematic Sampling

**Further Readings**

Burton, M., Della Croce, M., Masri, S. A., Bartholomew, M., & Yefremain, A. (2005). Sampling from the

United States Census Archives. *Field Methods, 17*(1), 102-118.
Ferguson, D. A. (2009). Name-based cluster sampling. *Sociological Methods & Research, 37*(4), 590–598.
Lam, L. T., & Yang, L. (2008). Duration of sleep and ADHD tendency among adolescents in China. *Journal of Attention Disorders, 11,* 437–444.
Pedulla, J. J., & Airasian, P. W. (1980). Sampling from samples: A comparison of strategies in longitudinal research. *Educational & Psychological Measurement, 40,* 807–813.

## Coefficient Alpha

Coefficient alpha, or *Cronbach's alpha,* is one way to quantify reliability and represents the proportion of observed score variance that is true score variance. Reliability is a property of a test that is derived from true scores, observed scores, and measurement error. Scores or values that are obtained from the measurement of some attribute or characteristic of a person (e.g., level of intelligence, preference for types of foods, spelling achievement, body length) are referred to as *observed scores.* In contrast, *true scores* are the scores one would obtain if these characteristics were measured without any random error. For example, every time you go to the doctor, the nurse measures your height. That is the observed height or observed "score" by that particular nurse. You return for another visit 6 months later, and another nurse measures your height. Again, that is an observed score. If you are an adult, it is expected that your true height has not changed in the 6 months since you last went to the doctor, but the two values might be different by .5 inch. When measuring the quantity of anything, whether it is a physical characteristic such as height or a psychological characteristic such as food preferences, spelling achievement, or level of intelligence, it is expected that the measurement will always be unreliable to some extent. That is, there is no perfectly reliable measure. Therefore, the observed score is the true score plus some amount of error, or an error score.

Measurement error can come from many different sources. For example, one nurse may have taken a more careful measurement of your height than the other nurse. Or you may have stood up straighter the first time you were measured. Measurement error for psychological attributes such as preferences, values, attitudes, achievement, and intelligence can also influence observed scores. For example, on the day of a spelling test, a child could have a cold that may negatively influence how well she would perform on the test. She may get a 70% on a test when she actually knew 80% of the material. That is, her observed score may be lower than her true score in spelling achievement. Thus, temporary factors such as physical health, emotional state of mind, guessing, outside distractions, misreading answers, or misrecording answers would artificially inflate or deflate the true scores for a characteristic. Characteristics of the test or the test administration can also create measurement error.

Ideally, test users would like to interpret individual's observed scores on a measure to reflect the person's true characteristic, whether it is physical (e.g., blood pressure, weight) or psychological (e.g., knowledge of world history, level of self-esteem). In order to evaluate the reliability of scores for any measure, one must estimate the extent to which individual differences are of function of the real or true score differences among respondents versus the extent to which they are a function of measurement error. A test that is considered reliable minimizes the measurement error so that error is not highly correlated with true score. That is, the relationship between the true score and observed score should be strong.

Given the assumptions of classical test theory, observed or empirical test scores can be used to estimate measurement error. There are several different ways to calculate empirical estimates of reliability. These include test–retest reliability, alternate- or parallel-forms reliability, and internal consistency reliability. To calculate test–retest reliability, respondents must take the same test or measurement twice (e.g., the same nurse measuring your height at two different points in time). Alternate-forms reliability requires the construction of two tests that measure the same set of true scores and have the same amount of error variance (this is theoretically possible, but difficult in a practical sense). Thus, the respondent completes both forms of the test in order to determine reliability of the measures. *Internal consistency reliability* is a practical alternative to

the test–retest and parallel-forms procedures because the respondents have to complete only one test at any one time. One form of internal consistency reliability is *split-half reliability,* in which the items for a measure are divided into two parts or subtests (e.g., odd- and even-numbered items), composite scores are computed for each subtest, and the two composite scores are correlated to provide an estimate of total test reliability. (This value is then adjusted by means of the Spearman–Brown prophecy formula.) Split-half reliability is not used very often, because it is difficult for the two halves of the test to meet the criteria of being "parallel." That is, how the test is divided is likely to lead to substantially different estimates of reliability.

The most widely used method of estimating reliability is coefficient alpha ($\alpha$), which is an estimate of internal consistency reliability. Lee Cronbach's often-cited article entitled "Coefficient Alpha and the Internal Structure of Tests" was published in 1951. This coefficient proved very useful for several reasons. First only one test administration was required rather than more than one, as in test–retest or parallel-forms estimates of reliability. Second, this formula could be applied to dichotomously scored items or polytomous items. Finally, it was easy to calculate, at a time before most people had access to computers, from the statistics learned in a basic statistics course.

Coefficient alpha is also know as the "raw" coefficient alpha. This method and other methods of determining internal consistency reliability (e.g., the generalized Spearman–Brown formula or the standardized alpha estimate) have at least two advantages over the lesser-used split-half method. First, they use more information about the test than the split-half method does. Imagine if a split-half reliability was computed, and then we randomly divided the items from the same sample into another set of split halves and recomputed, and kept doing this with all possible combinations of split-half estimates of reliability. Cronbach's alpha is mathematically equivalent to all possible split-half estimates, although it is not computed that way. Second, methods of calculating internal consistency estimates require fewer assumptions about the statistical properties of the individual items than does the split-half method of determining internal consistency reliability. Thus, internal consistency reliability is an estimate of how well the sum score on the items captures the true score on the entire domain from which the items are derived.

Internal consistency, such as measured by Cronbach's alpha, is a measure of the homogeneity of the items. When the various items of an instrument are measuring the same construct (e.g., depression, knowledge of subtraction), then scores on the items will tend to covary. That is, people will tend to score the same way across many items. The items on a test that has adequate or better internal consistency will be highly intecorrelated. Internal consistency reliability is the most appropriate type of reliability for assessing dynamic traits or traits that change over time, such as test anxiety or elated mood.

## Calculating Coefficient Alpha

The split-half method of determining internal consistency is based on the assumption that the two halves represent parallel subtests and that the correlation between the two halves produces the reliability estimate. For the "item-level" approach, such as the coefficient alpha, the logic of the split-half approach is taken further in that each item is viewed as a subtest. Thus, the association between items can be used to represent the reliability of the entire test. The item-level approach is a two-step process. In the first step, item-level statistics are calculated (item variances, interitem covariances, or interitem correlations). In the second step, the item-level information is entered into specialized equations to estimate the reliability of the complete test.

Below is the specialized formula for the calculation of coefficient alpha. Note that the first step is to determine the variance of scores on the complete test.

$$\alpha = \left(\frac{n}{n-1}\right) \times \left(\frac{SD_X^2 - \sum SD_i^2}{SD_X^2}\right),$$

where n equals the number of components (items or subtests), $SD_X^2$ is the variance of the observed total test scores, and $SD_i^2$ is the variance of component *i*.

An alternate way to calculate coefficient alpha is

$$\alpha = \frac{N - \bar{r}}{(\bar{v} + (N-1) - \bar{r})},$$

where $N$ equals the number of components, $\bar{v}$ is the average variance, and $\bar{r}$ is the average of all Pearson correlation coefficients between components.

The methods of calculating coefficient alpha as just described use raw scores (i.e., no transformation has been made to the item scores). There is also the standardized coefficient alpha described in some statistical packages as "Cronbach's Alpha Based on Standardized Items" (SPSS [an IBM company, formerly called PASW® Statistics] Reliability Analysis procedure) or "Cronbach Coefficient Alpha for Standardized Variables" (SAS). This standardized alpha estimate of reliability provides an estimate of reliability for an instrument in which scores on all items have been standardized to have equal means and standard deviations. The use of the raw score formula or the standardized score formula often produces roughly similar results.

## Optimizing Coefficient Alpha

There are four (at least) basic ways to influence the magnitude of Cronbach's coefficient alpha. The first has to with the characteristics of the sample (e.g., homogeneity vs. heterogeneity). Second and third, respectively, are the characteristics of the sample (e.g., size) and the number of items in the instrument. The final basic way occurs during the construction of the instrument.

Reliability is sample specific. A homogeneous sample with reduced true score variability may reduce the alpha coefficient. Therefore, if you administered an instrument measuring depression to only those patients recently hospitalized for depression, the reliability may be somewhat low because all participants in the sample have already been diagnosed with severe depression, and so there is not likely to be much variability in the scores on the measure. However, if one were to administer the same instrument to a general population, the larger variance in depression scores would be likely to produce higher reliability estimates. Therefore, to optimize Cronbach's coefficient alpha, it would be important to use a heterogeneous sample, which would have maximum true score variability. The sample size is also likely to influence the magnitude of coefficient alpha. This is because a larger size is more likely to produce a larger variance in the true scores. Typically, large numbers of subjects (typically in excess of 200) are needed to obtain generalizable reliability estimates.

The easiest way to increase coefficient alpha is by increasing the number of items. Therefore, if, for example, a researcher is developing a new measure of depression, he or she may want to begin with a large number of items to assess various aspects of depression (e.g., depressed mood, feeling fatigued, loss of interest in favorite activities). That is, the total variance becomes larger, relative to the sum of the variances of the items, as the number of items is increased. It can also be shown that when the interitem correlations are about the same, alpha approaches one as the number of items approaches infinity. However, it is also true that reliability is expected to be high even when the number of items is relatively small if the correlations among them are high. For example, a measure with 3 items whose average intercorrelation is .50 is expected to have a Cronbach's alpha coefficient of .75. This same alpha coefficient of .75 can be calculated from a measure composed of 9 items with an average intercorrelation among the 9 items of .25 and of 27 items when the average intercorrelation among them is .10.

Selecting "good" items during the construction of an instrument is another way to optimize the alpha coefficient. That is, scale developers typically want items that correlate highly with each other. Most statistical packages provide an item-total correlation as well as a calculation of the internal consistency reliability if any single item were removed. So one can choose to remove any items that reduce the internal consistency reliability coefficient.

## Interpreting Cronbach's Alpha Coefficient

A reliable test minimizes random measurement error so that error is not highly correlated with

the true scores. The relationship between the true score and the observed scores should be strong. A reliability coefficient is the proportion of the observed score variance that is true score variance. Thus, a coefficient alpha of .70 for a test means that 30% of the variance in scores is random and not meaningful. Rules of thumb exist for interpreting the size of coefficient alphas. Typically, a "high" reliability coefficient is considered to be .90 or above, "very good" is .80 to .89, and "good" or "adequate" is .70 to .79. Cronbach's alpha is a lower-bound estimate. That is, the actual reliability may be slightly higher. It is also considered to be the most accurate type of reliability estimate within the classical theory approach, along with the Kuder–Richardson 20, which is used only for dichotomous variables.

The interpretation of coefficient alpha and other types of reliability depends to some extent on just what is being measured. When tests are used to make important decisions about people, it would be essential to have high reliability (e.g., .90 or above). For example, individualized intelligence tests have high internal consistency. Often an intelligence test is used to make important final decisions. In contrast, lower reliability (e.g., .60 to .80) may be acceptable for looking at group differences in such personality characteristics as the level of extroversion.

It should be noted that the internal consistency approach applied through coefficient alpha assumes that the items, or subparts, of an instrument measure the same construct. Broadly speaking, that means the items are homogeneous. However, there is no general agreement about what the term *homogeneity* means and how it might be measured. Some authors interpret it to mean unidimensionality, or having only one factor. However, Cronbach did not limit alpha to an instrument with only one factor. In fact, he said in his 1951 article, "Alpha estimates the proportion of the test variance due to all common factors among items. That is, it reports how much the test score depends upon general and group rather than item specific factors" (p. 320). Cronbach's "general" factor is the first or most important factor, and alpha can be high even if there is no general factor underlying the relations among the items. This will happen if two or more "group" or common factors underlie the relations among the items. That is, an instrument with distinct sets of items or factors can still have an average intercorrelation among the items that is relatively large and would then result in a high alpha coefficient. Also, as discussed previously, even if the average intercorrelation among items is relatively small, the alpha can be high if the number of items is relatively large. Therefore, a high internal consistency estimate cannot serve as evidence of the homogeneity of a measure. However, a low internal consistency reliability coefficient does mean that the measure is not homogeneous because the items do not correlate well together.

It is important to keep several points in mind about reliability when designing or choosing an instrument. First, although reliability reflects variance due to true scores, it does not indicate what the true scores are measuring. If one uses a measure with a high internal consistency (.90), it may be that the instrument is measuring something different from what is postulated. For example, a personality instrument may be assessing social desirability and not the stated construct. Often the name of the instrument indicates the construct being tapped (e.g., the XYZ Scale of Altruism), but adequate reliability does not mean that the measure is assessing what it purports to measure (in this example, altruism). That is a validity argument.

*Karen D. Multon and Jill S. M. Coleman*

***See also*** Classical Test Theory; "Coefficient Alpha and the Internal Structure of Tests"; Correlation; Instrumentation; Internal Consistency Reliability; Pearson Product-Moment Correlation Coefficient; Reliability; Spearman–Brown Prophecy Formula; "Validity"

**Further Readings**

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and application. *Journal of Applied Psychology, 78*, 98–104.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Osburn, H. G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological Methods, 5,* 343–355.

## "Coefficient Alpha and the Internal Structure of Tests"

Lee Cronbach's 1951 *Psychometrika* article "Coefficient Alpha and the Internal Structure of Tests" established coefficient alpha as the preeminent estimate of internal consistency reliability. Cronbach demonstrated that coefficient alpha is the mean of all split-half reliability coefficients and discussed the manner in which coefficient alpha should be interpreted. Specifically, alpha estimates the correlation between two randomly parallel tests administered at the same time and drawn from a universe of items like those in the original test. Further, Cronbach showed that alpha does not require the assumption that items be unidimensional. In his reflections 50 years later, Cronbach described how coefficient alpha fits within *generalizability theory,* which may be employed to obtain more informative explanations of test score variance.

Concerns about the accuracy of test scores are commonly addressed by computing *reliability coefficients*. An *internal consistency reliability coefficient,* which may be obtained from a single test administration, estimates the consistency of scores on repeated test administrations taking place at the same time (i.e., no changes in examinees from one test to the next). *Split-half reliability coefficients,* which estimate internal consistency reliability, were established as a standard of practice for much of the early 20th century, but such coefficients are not unique because they depend on particular splits of items into half tests. Cronbach presented coefficient alpha as an alternative method for estimating internal consistency reliability. Alpha is computed as follows:

$$\frac{k}{k-1}\left(1-\frac{\sum_i s_i^2}{s_t^2}\right),$$

where $k$ is the number of items, $s_i^2$ is the variance of scores on item $i$, and $s_t^2$ is the variance of total test scores. As demonstrated by Cronbach, alpha is the mean of all possible split-half coefficients for a test. Alpha is generally applicable for studying measurement consistency whenever data include multiple observations of individuals (e.g., item scores, ratings from multiple judges, stability of performance over multiple trials). Cronbach showed that the well-known Kuder–Richardson formula 20 (KR-20), which preceded alpha, was a special case of alpha when items are scored dichotomously.

One sort of internal consistency reliability coefficient, the *coefficient of precision,* estimates the correlation between a test and a hypothetical replicated administration of the same test when no changes in the examinees have occurred. In contrast, Cronbach explained that alpha, which estimates the *coefficient of equivalence,* reflects the correlation between two different $k$-item tests randomly drawn (without replacement) from a universe of items like those in the test and administered simultaneously. Since the correlation of a test with itself would be higher than the correlation between different tests drawn randomly from a pool, alpha provides a lower bound for the coefficient of precision. Note that alpha (and other internal consistency reliability coefficients) provides no information about variation in test scores that could occur if repeated testings were separated in time. Thus, some have argued that such coefficients overstate reliability.

Cronbach dismissed the notion that alpha requires the assumption of item unidimensionality (i.e., all items measure the same aspect of individual differences). Instead, alpha provides an estimate (lower bound) of the proportion of variance in test scores attributable to all common factors accounting for item responses. Thus, alpha can reasonably be applied to tests typically administered in educational settings and that comprise items that call on several skills or aspects of understanding in different combinations across items. Coefficient alpha, then, climaxed 50 years of work on correlational conceptions of reliability begun by Charles Spearman.

In a 2004 article published posthumously, "My Current Thoughts on Coefficient Alpha and Successor Procedures," Cronbach expressed doubt that coefficient alpha was the best way of judging reliability. It covered only a small part of the range

of measurement uses, and consequently it should be viewed within a much larger system of reliability analysis, generalizability theory. Moreover, alpha focused attention on reliability coefficients when that attention should instead be cast on measurement error and the standard error of measurement.

For Cronbach, the extension of alpha (and classical test theory) came when Fisherian notions of experimental design and analysis of variance were put together with the idea that some "treatment" conditions could be considered random samples from a large universe, as alpha assumes about item sampling. Measurement data, then, could be collected in complex designs with multiple variables (e.g., items, occasions, and rater effects) and analyzed with random-effects analysis of variance models. The goal was not so much to estimate a reliability coefficient as to estimate the components of variance that arose from multiple variables and their interactions in order to account for observed score variance. This approach of partitioning effects into their variance components provides information as to the magnitude of each of the multiple sources of error and a standard error of measurement, as well as an "alpha-like" reliability coefficient for complex measurement designs. Moreover, the variance-component approach can provide the value of "alpha" expected by increasing or decreasing the number of items (or raters or occasions) like those in the test. In addition, the proportion of observed score variance attributable to variance in item difficulty (or, for example, rater stringency) may also be computed, which is especially important to contemporary testing programs that seek to determine whether examinees have achieved an absolute, rather than relative, level of proficiency. Once these possibilities were envisioned, coefficient alpha morphed into generalizability theory, with sophisticated analyses involving crossed and nested designs with random and fixed variables (*facets*) producing variance components for multiple measurement facets such as raters and testing occasions so as to provide a complex standard error of measurement.

By all accounts, coefficient alpha—Cronbach's alpha—has been and will continue to be the most popular method for estimating behavioral measurement reliability. As of 2004, the 1951 coefficient alpha article had been cited in more than 5,000 publications.

*Jeffrey T. Steedle and Richard J. Shavelson*

***See also*** Classical Test Theory; Generalizability Theory; Internal Consistency Reliability; KR-20; Reliability; Split-Half Reliability

**Further Readings**

Brennan, R. L. (2001). *Generalizability theory.* New York: Springer-Verlag.

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational & Psychological Measurement, 64*(3), 391–418.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (pp. 65–110). Westport, CT: Praeger.

Shavelson, R. J. (2004). Editor's preface to Lee J. Cronbach's "My Current Thoughts on Coefficient Alpha and Successor Procedures." *Educational & Psychological Measurement, 64*(3), 389–390.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage.

## Coefficient of Concordance

Proposed by Maurice G. Kendall and Bernard Babington Smith, Kendall's coefficient of concordance ($W$) is a measure of the agreement among several ($m$) quantitative or semiquantitative variables that are assessing a set of $n$ objects of interest. In the social sciences, the variables are often people, called *judges,* assessing different subjects or situations. In community ecology, they may be species whose abundances are used to assess habitat quality at study sites. In taxonomy, they may be characteristics measured over different species, biological populations, or individuals.

There is a close relationship between Milton Friedman's two-way analysis of variance without replication by ranks and Kendall's coefficient of concordance. They address hypotheses concerning the same data table, and they use the same $\chi^2$ statistic for testing. They differ only in the formulation of their respective null hypothesis. Consider Table 1, which contains illustrative data. In Friedman's test, the null hypothesis is that there is no

**Table 1** Illustrative Example: Ranked Relative Abundances of Four Soil Mite Species (Variables) at 10 Sites (Objects)

| | *Ranks (column-wise)* | | | | *Sum of Ranks* |
|---|---|---|---|---|---|
| | *Species 13* | *Species 14* | *Species 15* | *Species 23* | $R_i$ |
| Site 4 | 5 | 6 | 3 | 5 | 19.0 |
| Site 9 | 10 | 4 | 8 | 2 | 24.0 |
| Site 14 | 7 | 8 | 5 | 4 | 24.0 |
| Site 22 | 8 | 10 | 9 | 2 | 29.0 |
| Site 31 | 6 | 5 | 7 | 6 | 24.0 |
| Site 34 | 9 | 7 | 10 | 7 | 33.0 |
| Site 45 | 3 | 3 | 2 | 8 | 16.0 |
| Site 53 | 1.5 | 2 | 4 | 9 | 16.5 |
| Site 61 | 1.5 | 1 | 1 | 2 | 5.5 |
| Site 69 | 4 | 9 | 6 | 10 | 29.0 |

*Source:* Legendre, P. (2005) Species associations: The Kendall coefficient of concordance revisited. *Journal of Agricultural, Biological, & Environmental Statistics, 10*, 230. Reprinted with permission from the *Journal of Agricultural, Biological, & Environmental Statistics*. Copyright 2005 by the American Statistical Association. All rights reserved.

*Notes:* The ranks are computed columnwise with ties. Right-hand column: sum of the ranks for each site.

real difference among the $n$ objects (sites, rows of Table 1) because they pertain to the same statistical population. Under the null hypothesis, they should have received random ranks along the various variables, so that their sums of ranks should be approximately equal. Kendall's test focuses on the $m$ variables. If the null hypothesis of Friedman's test is true, this means that the variables have produced rankings of the objects that are independent of one another. This is the null hypothesis of Kendall's test.

## Computing Kendall's *W*

There are two ways of computing Kendall's $W$ statistic (first and second forms of Equations 1 and 2); they lead to the same result. $S$ or $S'$ is computed first from the row-marginal sums of ranks $R_i$ received by the objects:

$$S = \sum_{i=1}^{n} (R_i - \overline{R})^2 \text{ or } S' = \sum_{i=1}^{n} R_i^2 = SSR, \quad (1)$$

where $S$ is a sum-of-squares statistic over the row sums of ranks $R_i$, and $\overline{R}$ is the mean of the $R_i$ values. Following that, Kendall's $W$ statistic can be obtained from either of the following formulas:

$$W = \frac{12S}{m^2(n^3 - n) - mT}$$

or

$$W = \frac{12S' - 3m^2n(n+1)^2}{m^2(n^3 - n) - mT}, \quad (2)$$

where $n$ is the number of objects and $m$ is the number of variables. $T$ is a correction factor for tied ranks:

$$T = \sum_{k=1}^{g} (t_k^3 - t_k), \quad (3)$$

in which $t_k$ is the number of tied ranks in each ($k$) of $g$ groups of ties. The sum is computed over all groups of ties found in all $m$ variables of the data table. $T = 0$ when there are no tied values.

Kendall's $W$ is an estimate of the variance of the row sums of ranks $R_i$ divided by the maximum possible value the variance can take; this occurs when all variables are in total agreement. Hence $0 \leq W \leq 1$, 1 representing perfect concordance. To derive the formulas for $W$ (Equation 2), one has to know that when all variables are in perfect agreement, the sum of all sums of ranks in the data table (right-hand column of Table 1) is $mn(n+1)/2$ and that the sum of squares of the sums of all ranks is $m^2n(n+1)(2n+1)/6$ (without ties).

There is a close relationship between Charles Spearman's correlation coefficient $r_S$ and Kendall's $W$ statistic: $W$ can be directly calculated from the

mean ($\bar{r}_S$) of the pairwise Spearman correlations $r_S$ using the following relationship:

$$W = \frac{(m-1)\bar{r}_S + 1}{m}, \qquad (4)$$

where $m$ is the number of variables (judges) among which Spearman correlations are computed. Equation 4 is strictly true for untied observations only; for tied observations, ties are handled in a bivariate way in each Spearman $r_S$ coefficient whereas in Kendall's $W$ the correction for ties is computed in a single equation (Equation 3) for all variables. For two variables (judges) only, $W$ is simply a linear transformation of $r_S$: $W = (r_S + 1)/2$. In that case, a permutation test of $W$ for two variables is the exact equivalent of a permutation test of $r_S$ for the same variables.

The relationship described by Equation 4 clearly limits the domain of application of the coefficient of concordance to variables that are all meant to estimate the same general property of the objects: variables are considered concordant only if their Spearman correlations are positive. Two variables that give perfectly opposite ranks to a set of objects have a Spearman correlation of $-1$, hence $W = 0$ for these two variables (Equation 4); this is the lower bound of the coefficient of concordance. For two variables only, $r_S = 0$ gives $W = 0.5$. So coefficient $W$ applies well to rankings given by a panel of judges called in to assess overall performance in sports or quality of wines or food in restaurants, to rankings obtained from criteria used in quality tests of appliances or services by consumer organizations, and so forth. It does not apply, however, to variables used in multivariate analysis in which negative as well as positive relationships are informative. Jerrold H. Zar, for example, uses wing length, tail length, and bill length of birds to illustrate the use of the coefficient of concordance. These data are appropriate for $W$ because they are all indirect measures of a common property, the size of the birds.

In ecological applications, one can use the abundances of various species as indicators of the good or bad environmental quality of the study sites. If a group of species is used to produce a global index of the overall quality (good or bad) of the environment at the study sites, only the species that are significantly associated and positively correlated to one another should be included in the index, because different groups of species may be associated to different environmental conditions.

## Testing the Significance of $W$

Friedman's chi-square statistic is obtained from $W$ by the formula

$$\chi^2 = m(n-1)W. \qquad (5)$$

This quantity is asymptotically distributed like chi-square with $\nu = (n-1)$ degrees of freedom; it can be used to test $W$ for significance. According to Kendall and Babington Smith, this approach is satisfactory only for moderately large values of $m$ and $n$.

Sidney Siegel and N. John Castellan Jr. recommend the use of a table of critical values for $W$ when $n \leq 7$ and $m \leq 20$; otherwise, they recommend testing the chi-square statistic (Equation 5) using the chi-square distribution. Their table of critical values of $W$ for small $n$ and $m$ is derived from a table of critical values of $S$ assembled by Friedman using the $z$ test of Kendall and Babington Smith and reproduced in Kendall's classic monograph, *Rank Correlation Methods*. Using numerical simulations, Pierre Legendre compared results of the classical chi-square test of the chi-square statistic (Equation 5) to the permutation test that Siegel and Castellan also recommend for small samples (small $n$). The simulation results showed that the classical chi-square test was too conservative for any sample size ($n$) when the number of variables $m$ was smaller than 20; the test had rejection rates well below the significance level, so it remained valid. The classical chi-square test had a correct level of Type I error (rejecting a null hypothesis that is true) for 20 variables and more. The permutation test had a correct rate of Type I error for all values of $m$ and $n$. The power of the permutation test was higher than that of the classical chi-square test because of the differences in rates of Type I error between the two tests. The differences in power disappeared asymptotically as the number of variables increased.

An alternative approach is to compute the following $F$ statistic:

$$F = (m-1)W/(1-W), \qquad (6)$$

which is asymptotically distributed like $F$ with $\nu_1 = n - 1 - (2/m)$ and $\nu_2 = \nu_1(m-1)$ degrees

of freedom. Kendall and Babington Smith described this approach using a Fisher $z$ transformation of the $F$ statistic, $z = 0.5 \log_e(F)$. They recommended it for testing $W$ for moderate values of $m$ and $n$. Numerical simulations show, however, that this $F$ statistic has correct levels of Type I error for any value of $n$ and $m$.

In permutation tests of Kendall's $W$, the objects are the permutable units under the null hypothesis (the objects are sites in Table 1). For the global test of significance, the rank values in all variables are permuted at random, independently from variable to variable because the null hypothesis is the independence of the rankings produced by all variables. The alternative hypothesis is that at least one of the variables is concordant with one, or with some, of the other variables. Actually, for permutation testing, the four statistics $SSR$ (Equation 1), $W$ (Equation 2), $\chi^2$ (Equation 5), and $F$ (Equation 6) are monotonic to one another since $n$ and $m$, as well as $T$, are constant within a given permutation test; thus they are equivalent statistics for testing, producing the same permutational probabilities. The test is one-tailed because it recognizes only positive associations between vectors of ranks. This may be seen if one considers two vectors with exactly opposite rankings: They produce a Spearman statistic of $-1$, hence a value of zero for $W$ (Equation 4).

Many of the problems subjected to Kendall's concordance analysis involve fewer than 20 variables. The chi-square test should be avoided in these cases. The $F$ test (Equation 6), as well as the permutation test, can safely be used with all values of $m$ and $n$.

## Contributions of Individual Variables to Kendall's Concordance

The overall permutation test of $W$ suggests a way of testing a posteriori the significance of the contributions of individual variables to the overall concordance to determine which of the individual variables are concordant with one or several other variables in the group. There is interest in several fields in identifying discordant variables or judges. This includes all fields that use panels of judges to assess the overall quality of the objects or subjects under study (sports, law, consumer protection, etc.). In other types of studies, scientists are interested in identifying variables that agree in their estimation of a common property of the objects. This is the case in environmental studies in which scientists are interested in identifying groups of concordant species that are indicators of some property of the environment and can be combined into indices of its quality, in particular in situations of pollution or contamination.

The contribution of individual variables to the $W$ statistic can be assessed by a permutation test proposed by Legendre. The null hypothesis is the monotonic independence of the variable subjected to the test, with respect to all the other variables in the group under study. The alternative hypothesis is that this variable is concordant with other variables in the set under study, having similar rankings of values (one-tailed test). The statistic $W$ can be used directly in a posteriori tests. Contrary to the global test, only the variable under test is permuted here. If that variable has values that are monotonically independent of the other variables, permuting its values at random should have little influence on the $W$ statistic. If, on the contrary, it is concordant with one or several other variables, permuting its values at random should break the concordance and induce a noticeable decrease on $W$.

Two specific partial concordance statistics can also be used in a posteriori tests. The first one is the mean, $\overline{r}_j$, of the pairwise Spearman correlations between variable $j$ under test and all the other variables. The second statistic, $W_j$, is obtained by applying Equation 4 to $\overline{r}_j$ instead of $\overline{r}$, with $m$ the number of variables in the group. These two statistics are shown in Table 2 for the example data; $\overline{r}_j$ and $W_j$ are monotonic to each other because $m$ is constant in a given permutation test. Within a given a posteriori test, $W$ is also monotonic to $W_j$ because only the values related to variable $j$ are permuted when testing variable $j$. These three statistics are thus equivalent for a posteriori permutation tests, producing the same permutational probabilities. Like $\overline{r}_j$, $W_j$ can take negative values; this is not the case of $W$.

There are advantages to performing a single a posteriori test for variable $j$ instead of $(m-1)$ tests of the Spearman correlation coefficients between variable $j$ and all the other variables: The tests of the $(m-1)$ correlation coefficients would

**Table 2** Results of (a) the Overall and (b) the A Posteriori Tests of Concordance Among the Four Species of Table 1; (c) Overall and (d) A Posteriori Tests of Concordance Among Three Species

(a) Overall test of $W$ statistic, four species. $H_0$ : The four species are not concordant with one another.

| | | |
|---|---|---|
| Kendall's $W$ = | 0.44160 | Permutational $p$ value = .0448* |
| $F$ statistic = | 2.37252 | $F$ distribution $p$ value = .0440* |
| Friedman's chi-square = | 15.89771 | Chi-square distribution $p$ value = .0690 |

(b) A posteriori tests, four species. $H_0$ : This species is not concordant with the other three.

| | $\bar{r}_j$ | $W_j$ | $p$ Value | Corrected $p$ | Decision at $\alpha = 5\%$ |
|---|---|---|---|---|---|
| Species 13 | 0.32657 | 0.49493 | .0766 | .1532 | Do not reject $H_0$ |
| Species 14 | 0.39655 | 0.54741 | .0240 | .0720 | Do not reject $H_0$ |
| Species 15 | 0.45704 | 0.59278 | .0051 | .0204* | Reject $H_0$ |
| Species 23 | −0.16813 | 0.12391 | .7070 | .7070 | Do not reject $H_0$ |

(c) Overall test of $W$ statistic, three species. $H_0$ : The three species are not concordant with one another.

| | | |
|---|---|---|
| Kendall's $W$ = | 0.78273 | Permutational $p$ value = .0005* |
| $F$ statistic = | 7.20497 | $F$ distribution $p$ value = .0003* |
| Friedman's chi-square = | 21.13360 | Chi-square distribution $p$ value = .0121* |

(d) A posteriori tests, three species. $H_0$ : This species is not concordant with the other two.

| | $\bar{r}_j$ | $W_j$ | $p$ Value | Corrected $p$ | Decision at $\alpha = 5\%$ |
|---|---|---|---|---|---|
| Species 13 | 0.69909 | 0.79939 | .0040 | .0120* | Reject $H_0$ |
| Species 14 | 0.59176 | 0.72784 | .0290 | .0290* | Reject $H_0$ |
| Species 15 | 0.73158 | 0.82105 | .0050 | .0120* | Reject $H_0$ |

*Source:* (a) and (b): Adapted from Legendre, P. (2005). Species associations: The Kendall coefficient of concordance revisited. *Journal of Agricultural, Biological, and Environmental Statistics, 10*, 233. Reprinted with permission from the *Journal of Agricultural, Biological and Environmental Statistics*. Copyright 2005 by the American Statistical Association. All rights reserved.

*Notes:* $\bar{r}_j$ = mean of the Spearman correlations with the other species; $W_j$ = partial concordance per species; $p$ value = permutational probability (9,999 random permutations); corrected $p$ = Holm-corrected $p$ value. * = Reject $H_0$ at $\alpha = .05$.

have to be corrected for multiple testing, and they could provide discordant information; a single test of the contribution of variable $j$ to the $W$ statistic has greater power and provides a single, clearer answer. In order to preserve a correct or approximately correct experimentwise error rate, the probabilities of the a posteriori tests computed for all species in a group should be adjusted for multiple testing.

A posteriori tests are useful for identifying the variables that are not concordant with the others, as in the examples, but they do not tell us whether there are one or several groups of congruent variables among those for which the null hypothesis of independence is rejected. This information can be obtained by computing Spearman correlations among the variables and clustering them into groups of variables that are significantly and positively correlated.

The example data are analyzed in Table 2. The overall permutational test of the $W$ statistic is significant at $\alpha = 5\%$, but marginally (Table 2a). The cause appears when examining the a posteriori tests in Table 2b: Species 23 has a negative mean correlation with the three other species in the group ($\bar{r}_j = -.168$). This indicates that Species 23 does not belong in that group. Were we analyzing a large group of variables, we could look at the next partition in an agglomerative clustering dendrogram, or the next $K$-means partition, and proceed to the overall and a posteriori tests for the members of these new groups. In the present illustrative example, Species 23 clearly differs from the other three species. We can now test Species 13, 14, and 15 as a group. Table 2c shows that this group has a highly significant concordance, and all individual species contribute significantly to the overall concordance of their group (Table 2d).

In Table 2a and 2c, the *F* test results are concordant with the permutation test results, but due to small *m* and *n,* the chi-square test lacks power.

### Discussion

The Kendall coefficient of concordance can be used to assess the degree to which a group of variables provides a common ranking for a set of objects. It should be used only to obtain a statement about variables that are all meant to measure the same general property of the objects. It should not be used to analyze sets of variables in which the negative and positive correlations have equal importance for interpretation. When the null hypothesis is rejected, one cannot conclude that all variables are concordant with one another, as shown in Table 2 (a) and (b); only that at least one variable is concordant with one or some of the others.

The partial concordance coefficients and a posteriori tests of significance are essential complements of the overall test of concordance. In several fields, there is interest in identifying discordant variables; this is the case in all fields that use panels of judges to assess the overall quality of the objects under study (e.g., sports, law, consumer protection). In other applications, one is interested in using the sum of ranks, or the sum of values, provided by several variables or judges, to create an overall indicator of the response of the objects under study. It is advisable to look for one or several groups of variables that rank the objects broadly in the same way, using clustering, and then carry out a posteriori tests on the putative members of each group. Only then can their values or ranks be pooled into an overall index.

*Pierre Legendre*

***See also*** Friedman Test; Holm's Sequential Bonferroni Procedure; Spearman Rank Order Correlation

#### Further Readings

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association, 32,* 675–701.

Friedman, M. (1940). A comparison of alternative tests of significance for the problem of *m* rankings. *Annals of Mathematical Statistics, 11,* 86–92.

Kendall, M. G. (1948). *Rank correlation methods* (1st ed.). London: Charles Griffith.

Kendall, M. G., & Babington Smith, B. (1939). The problem of *m* rankings. *Annals of Mathematical Statistics, 10,* 275–287.

Legendre, P. (2005). Species associations: The Kendall coefficient of concordance revisited. *Journal of Agricultural, Biological, & Environmental Statistics, 10,* 226–245.

Zar, J. H. (1999). *Biostatistical analysis* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

## Coefficient of Variation

The coefficient of variation measures the variability of a series of numbers independent of the unit of measurement used for these numbers. In order to do so, the coefficient of variation eliminates the unit of measurement of the standard deviation of a series of numbers by dividing the standard deviation by the mean of these numbers. The coefficient of variation can be used to compare distributions obtained with different units, such as the variability of the weights of newborns (measured in grams) with the size of adults (measured in centimeters). The coefficient of variation is meaningful only for measurements with a real zero (i.e., "ratio scales") because the mean is meaningful (i.e., unique) only for these scales. So, for example, it would be meaningless to compute the coefficient of variation of the temperature measured in degrees Fahrenheit, because changing the measurement to degrees Celsius will not change the temperature but will change the value of the coefficient of variation (because the value of zero for Celsius is 32 for Fahrenheit, and therefore the mean of the temperature will change from one scale to the other). In addition, the values of the measurement used to compute the coefficient of variation are assumed to be always positive or null. The coefficient of variation is primarily a descriptive statistic, but it is amenable to statistical inferences such as null hypothesis testing or confidence intervals. Standard procedures are often very dependent on the normality assumption,

and current work is exploring alternative procedures that are less dependent on this normality assumption.

## Definition and Notation

The coefficient of variation, denoted $C_v$ (or occasionally $V$), eliminates the unit of measurement from the standard deviation of a series of numbers by dividing it by the mean of this series of numbers. Formally, if, for a series of $N$ numbers, the standard deviation and the mean are denoted respectively by $S$ and $M$, the coefficient of variation is computed as

$$C_v = \frac{S}{M}. \tag{1}$$

Often the coefficient of variation is expressed as a percentage, which corresponds to the following formula for the coefficient of variation:

$$C_v = \frac{S \times 100}{M}. \tag{2}$$

This last formula can be potentially misleading because, as shown later, the value of the coefficient of variation can exceed 1 and therefore would create percentages larger than 100. In that case, Formula 1, which expresses $C_v$ as a ratio rather than a percentage, should be used.

## Range

In a finite sample of $N$ nonnegative numbers with a real zero, the coefficient of variation can take a value between 0 and $\sqrt{N-1}$ (the maximum value of $C_v$ is reached when all values but one are equal to zero).

### *Estimation of a Population Coefficient of Variation*

The coefficient of variation computed on a sample is a *biased* estimate of the population coefficient of variation denoted $\gamma_v$ An unbiased estimate of the population coefficient of variation, denoted $\hat{C}_v$, is computed as

$$\hat{C}_v = \left(1 + \frac{1}{4N}\right) C_v \tag{3}$$

(where $N$ is the sample size).

## Testing the Coefficient of Variation

When the coefficient of variation is computed on a sample drawn from a normal population, its standard error, denoted $\sigma_{C_v}$, is known and is equal to

$$\sigma_{C_v} = \frac{\gamma v}{\sqrt{2N}} \tag{4}$$

When $\gamma_v$ is not known (which is, in general, the case), $\sigma_{C_v}$ can be estimated by replacing $\gamma_v$ by its estimation from the sample. Either $C_v$ or $\hat{C}_v$ can be used for this purpose ($\hat{C}_v$ being preferable because it is a better estimate). So $\sigma_{C_v}$ can be estimated as

$$S_{C_v} = \frac{C_v}{\sqrt{2N}} \text{ or } \hat{S}_{C_v} = \frac{\hat{C}_v}{\sqrt{2N}}. \tag{5}$$

Therefore, under the assumption of normality, the statistic

$$t_{C_v} = \frac{C_v - \gamma_v}{S_{C_v}} \tag{6}$$

follows a Student distribution with $v = N - 1$ degrees of freedom. It should be stressed that this test is very sensitive to the normality assumption. Work is still being done to minimize the effect of this assumption.

If Equation (6) is rewritten, confidence intervals can be computed as

$$C_v \pm t_{\alpha,v} S_{C_v} \tag{7}$$

**Table I** Example for the Coefficient of Variation

| *Saleperson* | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* | *9* | *10* |
|---|---|---|---|---|---|---|---|---|---|---|
| Commission | 152 | 155 | 164 | 164 | 182 | 221 | 233 | 236 | 245 | 248 |

*Note:* Daily commission (in dollars) of 10 salespersons.

(with $t_{\alpha,\nu}$ being the critical value of Student's $t$ for the chosen $\alpha$ level and for $\nu = N - 1$ degrees of freedom). Again, because $\hat{C}_\nu$ is a better estimation of $\gamma_\nu$ than $C_\nu$ is, it makes sense to use $\hat{C}_\nu$ rather than $C_\nu$.

### Example

Table 1 lists the daily commission in dollars of 10 car salespersons. The mean commission is equal to \$200, with a standard deviation of \$40.

This gives a value of the coefficient of variation of

$$C_\nu = \frac{S}{M} = \frac{40}{200} = 0.200, \tag{8}$$

which corresponds to a population estimate of

$$\hat{C}_\nu = \left(1 + \frac{1}{4N}\right) C_\nu = \left(1 + \frac{1}{4 \times 10}\right) \times 0.200 = 0.205. \tag{9}$$

The standard error of the coefficient of variation is estimated as

$$S_{C_\nu} = \frac{C_\nu}{\sqrt{2N}} = \frac{0.200}{\sqrt{2 \times 10}} = 0.0447 \tag{10}$$

(the value of $\hat{S}_{C_\nu}$ is equal to 0.0458).

A $t$ criterion testing the hypothesis that the population value of the coefficient of variation is equal to zero is equal to

$$t_{C_\nu} = \frac{C_\nu - \gamma_\nu}{S_{C_\nu}} = \frac{0.2000}{0.0447} = 4.47. \tag{11}$$

This value of $t_{C_\nu} = 4.47$ is larger than the critical value of $t_{\alpha,\nu} = 2.26$ (which is the critical value of a Student's $t$ distribution for $\alpha = .05$ and $\nu = .9$ degrees of freedom). Therefore, we can reject the null hypothesis and conclude that $\gamma_\nu$ is larger than zero. A 95% corresponding confidence interval gives the values of

$$C_\nu \pm t_{\alpha,\nu} S_{C_\nu} = 0.2000 \pm 2.26 \times 0.0447 = 0.200 \pm 0.1011 \tag{12}$$

and therefore we conclude that there is a probability of .95 that the value of $\gamma_\nu$ lies in the interval [0.0989 to 0.3011].

*Hervé Abdi*

***See also*** Mean; Standard Deviation; Variability, Measure of; Variance

#### Further Readings


Abdi, H., Edelman, B., Valentin, D., & Dowling, W. J. (2009). *Experimental design and analysis for psychology.* Oxford, UK: Oxford University Press.

Curto, J. D., & Pinto, J. C. (2009). The coefficient of variation asymptotic distribution in the case of non-*iid* random variables. *Journal of Applied Statistics, 36,* 21–32.

Nairy, K. S., & Rao, K. N. (2003). Tests of coefficients of variation of normal populations. *Commnunications in Statistics Simulation and Computation, 32,* 641–661.

Martin, J. D., & Gray, L. N. (1971). Measurement of relative variation: Sociological examples. *American Sociological Review, 36,* 496–502.

Sokal, R. R., & Rohlf, F. J. (1995). *Biometry* (3rd ed.). New York: Freeman.


## Coefficients of Correlation, Alienation, and Determination

The coefficient of correlation evaluates the similarity of two sets of measurements (i.e., two dependent variables) obtained on the same observations. The coefficient of correlation indicates the amount of information common to the two variables. This coefficient takes values between $-1$ and $+1$ (inclusive). A value of $+1$ shows that the two series of measurements are measuring the same thing. A value of $-1$ indicates that the two measurements are measuring the same thing, but one measurement varies inversely to the other. A value of 0 indicates that the two series of measurements have nothing in common. It is important to note that the coefficient of correlation measures only

the *linear* relationship between two variables and that its value is very sensitive to outliers.

The squared correlation gives the proportion of common variance between two variables and is also called the *coefficient of determination*. Subtracting the coefficient of determination from unity gives the proportion of variance not shared between two variables. This quantity is called the *coefficient of alienation*.

The significance of the coefficient of correlation can be tested with an $F$ or a $t$ test. This entry presents three different approaches that can be used to obtain $p$ values: (1) the *classical* approach, which relies on Fisher's $F$ distributions; (2) the *Monte Carlo* approach, which relies on computer simulations to derive empirical approximations of sampling distributions; and (3) the *nonparametric permutation* (also known as *randomization*) test, which evaluates the likelihood of the actual data against the set of all possible configurations of these data. In addition to $p$ values, confidence intervals can be computed using Fisher's $Z$ transform or the more modern, computationally based, and nonparametric *Efron's bootstrap*.

Note that the coefficient of correlation always overestimates the intensity of the correlation in the population and needs to be "corrected" in order to provide a better estimation. The corrected value is called *shrunken* or *adjusted*.

## Notations and Definition

Suppose we have $S$ observations, and for each observation $s$, we have two measurements, denoted $W_s$ and $Y_s$, with respective means denoted $M_w$ and $M_y$. For each observation, we define the cross-product as the product of the deviations of each variable from its mean. The sum of these cross-products, denoted $SCP_{wy}$, is computed as

$$SCP_{WY} = \sum_{s}^{S} (W_s - M_W)(Y_s - M_Y). \quad (1)$$

The sum of the cross-products reflects the association between the variables. When the deviations have the same sign, they indicate a positive relationship, and when they have different signs, they indicate a negative relationship.

The average value of the $SCP_{WY}$ is called the covariance (just like the variance, the covariance can be computed by dividing by $S$ or by $S-1$):

$$\text{cov}_{WY} = \frac{SCP}{\text{Number of Observations}} = \frac{SCP}{S}. \quad (2)$$

The covariance reflects the association between the variables, but it is expressed in the original units of measurement. In order to eliminate the units, the covariance is normalized by division by the standard deviation of each variable. This defines the coefficient of correlation, denoted $r_{W.Y}$, which is equal to

$$r_{W.Y} = \frac{\text{cov}_{WY}}{\sigma_W \sigma_Y}. \quad (3)$$

Rewriting the previous formula gives a more practical formula:

$$r_{W.Y} = \frac{SCP_{WY}}{\sqrt{SS_W SS_Y}}. \quad (4)$$

where $SCP$ is the sum of the cross-product and $SS_W$ and $SS_Y$ are the sum of squares of $W$ and $Y$, respectively.

## Correlation Computation: An Example

The computation for the coefficient of correlation is illustrated with the following data, describing the values of $W$ and $Y$ for $S = 6$ subjects:

$$W_1 = 1, W_2 = 3, W_3 = 4, W_4 = 4, W_5 = 5, W_6 = 7$$
$$Y_1 = 16, Y_2 = 10, Y_3 = 12, Y_4 = 4, Y_5 = 8, Y_6 = 10.$$

*Step 1*. Compute the sum of the cross-products. First compute the means of $W$ and $Y$:

$$M_W = \frac{1}{S}\sum_{s=1}^{S} W_s = \frac{24}{6} = 4 \text{ and}$$
$$M_Y = \frac{1}{S}\sum_{s=1}^{S} Y_s = \frac{60}{6} = 10.$$

The sum of the cross-products is then equal to

$$
\begin{aligned}
SCP_{WY} &= \sum_{s}^{S} (Y_s - M_Y)(W_s - M_W) \\
&= (16-10)(1-4) \\
&\quad + (10-10)(3-4) \\
&\quad + (12-10)(4-4) \\
&\quad + (4-10)(4-4) \\
&\quad + (8-10)(5-4) \\
&\quad + (10-10)(7-4) \\
&= (6\times -3)(0\times -1) \\
&\quad + (2\times 0)(-6\times 0) \\
&\quad + (-2\times 1) + (0\times 3) \\
&= -18 + 0 + 0 + 0 - 2 + 0 \\
&= -20.
\end{aligned} \tag{5}
$$

*Step 2*. Compute the sums of squares. The sum of squares of $W_s$ is obtained as

$$
\begin{aligned}
SS_W &= \sum_{s=1}^{S} (W_s - M_W)^2 \\
&= (1-4)^2 + (3-4)^2 + (4-4)^2 \\
&\quad + (4-4)^2 + (5-4)^2 + (7-4)^2 \\
&= (-3)^2 + (-1)^2 + 0^2 + 0^2 \\
&\quad + 1^2 + 3^2 \\
&= 9 + 1 + 0 + 0 + 1 + 9 \\
&= -18 + 0 + 0 + 0 - 2 + 0 \\
&= 20.
\end{aligned} \tag{6}
$$

The sum of squares of $Y_s$ is

$$
\begin{aligned}
SS_Y &= \sum_{s=1}^{S} (Y_s - M_Y)^2 \\
&= (16-10)^2 + (10-10)^2 \\
&\quad + (12-10)^2 + (4-10)^2 + (8-10)^2 \\
&\quad + (10-10)^2 \\
&= 6^2 + 0^2 + 2^2 + (-6)^2 + (-2)^2 + 0^2 \\
&= 36 + 0 + 4 + 36 + 4 + 0 \\
&= 80.
\end{aligned} \tag{7}
$$

*Step 3*. Compute $r_{W.Y}$. The coefficient of correlation between $W$ and $Y$ is equal to

$$
\begin{aligned}
r_{W.Y} &= \frac{\sum_{s} (Y_s - M_Y)(W_s - M_W)}{\sqrt{SS_Y \times SS_W}} = \frac{SCP_{WY}}{\sqrt{SS_W SS_Y}} \\
&= \frac{-20}{\sqrt{80\times 20}} = \frac{-20}{\sqrt{1600}} = -\frac{20}{40} \\
&= -.5.
\end{aligned} \tag{8}
$$

This value of $r = -.5$ can be interpreted as an indication of a negative linear relationship between $W$ and $Y$.

## Properties of the Coefficient of Correlation

The coefficient of correlation is a number *without unit*. This occurs because dividing the units of the numerator by the same units in the denominator eliminates the units. Hence, the coefficient of correlation can be used to compare different studies performed using different variables.

The magnitude of the coefficient of correlation is always smaller than or equal to 1. This happens because the numerator of the coefficient of correlation (see Equation 4) is always smaller than or equal to its denominator (this property follows from the *Cauchy–Schwartz inequality*). A coefficient of correlation that is equal to $+1$ or $-1$ indicates that the plot of the observations will have all observations positioned on a line.

The squared coefficient of correlation gives the *proportion of common variance* between two variables. It is also called the *coefficient of determination*. In our example, the coefficient of determination is equal to $r^2_{WY} = .25$. The proportion of variance not shared between the variables is called the *coefficient of alienation*, and for our example, it is equal to $1 - r^2_{WY} = .75$.

## Interpreting Correlation

### *Linear and Nonlinear Relationship*

The coefficient of correlation measures only *linear* relationships between two variables and will miss *nonlinear* relationships. For example, Figure 1 displays a perfect nonlinear relationship between two variables (i.e., the data show a U-shaped relationship with $Y$ being proportional to the square of $W$), but the coefficient of correlation is equal to 0.

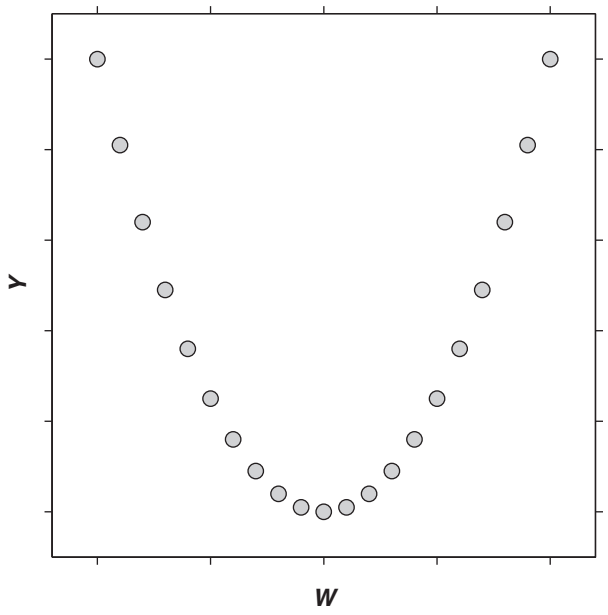



**Figure 1** A Perfect Nonlinear Relationship With a 0 Correlation ($r_{W.Y} = 0$)

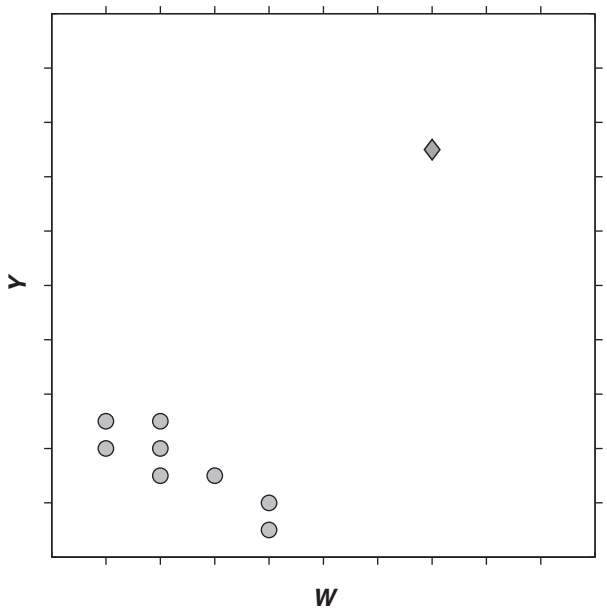



**Figure 2** The Dangerous Effect of Outliers on the Value of the Coefficient of Correlation

*Notes:* The correlation of the set of points represented by the circles is equal to $-.87$. When the point represented by the diamond is added to the set, the correlation is now equal to $+.61$, which shows that an outlier can determine the value of the coefficient of correlation.

### *Effect of Outliers*

Observations far from the center of the distribution contribute a lot to the sum of the cross-products. In fact, as illustrated in Figure 2, a single extremely deviant observation (often called an *outlier*) can dramatically influence the value of *r*.

### *Geometric Interpretation*

Each set of observations can also be seen as a *vector* in an *S* dimensional space (one dimension per observation). Within this framework, the correlation is equal to the *cosine* of the angle between the two vectors after they have been centered by subtracting their respective mean. For example, a coefficient of correlation of $r = -.50$ corresponds to a 150-degree angle. A coefficient of correlation of 0 corresponds to a right angle, and therefore two uncorrelated variables are called *orthogonal* (which is derived from the Greek word for right angle).

### *Correlation and Causation*

The fact that two variables are correlated does not mean that one variable causes the other one: *Correlation is not causation*. For example, in France, the number of Catholic churches in a city, as well as the number of schools, is highly correlated with the number of cases of cirrhosis of the liver, the number of teenage pregnancies, and the number of violent deaths. Does this mean that churches and schools are sources of vice and that newborns are murderers? Here, in fact, the observed correlation is due to a third variable, namely the size of the cities: the larger a city, the larger the number of churches, schools, alcoholics, and so forth. In this example, the correlation between number of churches or schools and alcoholics is called a *spurious* correlation because it reflects only their mutual correlation with a third variable (i.e., size of the city).

## Testing the Significance of *r*

A null hypothesis test for *r* can be performed using an *F* statistic obtained as

$$F = \frac{r^2}{1 - r^2} \times (S - 2). \tag{9}$$

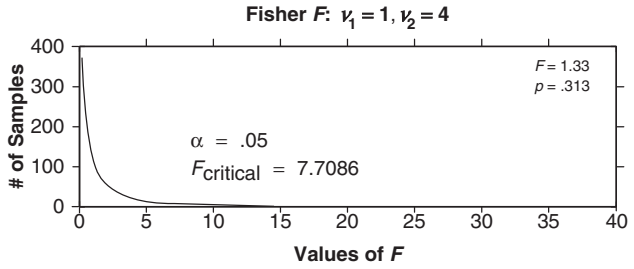



**Figure 3** The Fisher Distribution for $\nu_1 = 1$ and $\nu_2 = 4$, Along With $\alpha = .05$

*Note:* Critical value of $F = 7.7086$.

For our example, we find that

$$F = \frac{.25}{1 - .25} \times (6 - 2) = \frac{.25}{.75} \times 4 = \frac{1}{3} \times 4 = \frac{4}{3} = 1.33.$$

In order to perform a statistical test, the next step is to evaluate the sampling distribution of the *F*. This sampling distribution provides the probability of finding any given value of the *F* criterion (i.e., the *p* value) under the null hypothesis (i.e., when there is *no* correlation between the variables). If this *p* value is smaller than the chosen level (e.g., .05 or .01), then the null hypothesis can be rejected, and *r* is considered *significant*. The problem of finding the *p* value can be addressed in three ways: (1) the classical approach, which uses Fisher's *F* distributions; (2) the Monte Carlo approach, which generates empirical probability distributions; and (3) the (nonparametric) permutation test, which evaluates the likelihood of the actual configuration of results among all other possible configurations of results.

### *Classical Approach*

In order to analytically derive the sampling distribution of *F*, several assumptions need to be made: (a) the error of measurement is added to the true measure; (b) the error is independent of the measure; and (c) the mean error is normally distributed, has a mean of zero, and has a variance of $\sigma_e^2$. When theses assumptions hold and when the null hypothesis is true, the *F* statistic is distributed as a Fisher's *F* with $\nu_1 = 1$ and $\nu_2 = S - 2$ degrees of freedom. (Incidentally, an equivalent test can be performed using $t = \sqrt{F}$, which is distributed under $H_0$ as a Student's distribution with $\nu = S - 2$ degrees of freedom).

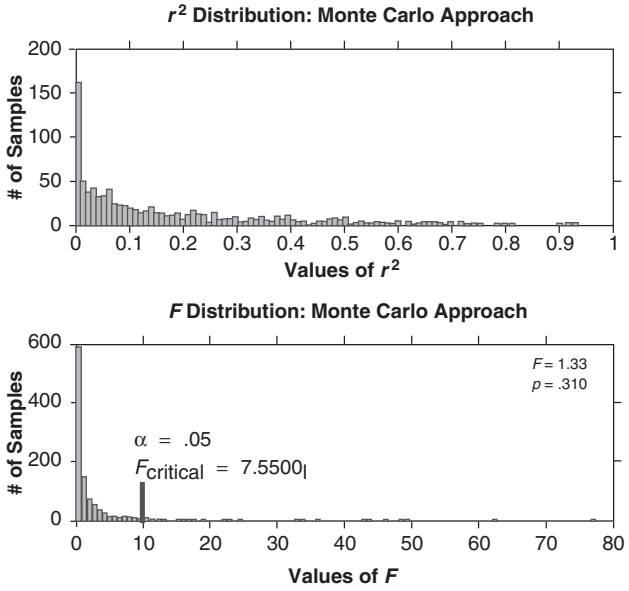



**Figure 4** Histogram of Values of $r^2$ and *F* Computed From 1,000 Random Samples When the Null Hypothesis Is True

*Notes:* The histograms show the *empirical* distribution of *F* and $r^2$ under the null hypothesis.

For our example, the Fisher distribution shown in Figure 3 has $\nu_1 = 1$ and $\nu_2 = S - 2 = 6 - 2 = 4$ and gives the sampling distribution of *F*. The use of this distribution will show that the probability of finding a value of $F = 1.33$ under $H_0$ is equal to $p \approx .313$ (most statistical packages will routinely provide this value). Such a *p* value does not lead to rejecting $H_0$ at the usual level of $\alpha = .05$ or $\alpha = .01$. An equivalent way of performing a test uses critical values that correspond to values of *F* whose *p* value is equal to a given $\alpha$ level. For our example, the critical value (found in tables available in most standard textbooks) for $\alpha = .05$ is equal to $F(1, 4) = 7.7086$. Any *F* with a value larger than the critical value leads to rejection of the null hypothesis at the chosen $\alpha$ level, whereas an *F* value smaller than the critical value leads one to fail to reject the null hypothesis. For our example, because $F = 1.33$ is smaller than the critical value of 7.7086, we cannot reject the null hypothesis.

### *Monte Carlo Approach*

A modern alternative to the analytical derivation of the sampling distribution is to *empirically* obtain the sampling distribution of $F$ when the null hypothesis is true. This approach is often called a *Monte Carlo* approach.

With the Monte Carlo approach, we generate a large number of random samples of observations (e.g., 1,000 or 10,000) and compute $r$ and $F$ for each sample. In order to generate these samples, we need to specify the shape of the population from which these samples are obtained. Let us use a normal distribution (this makes the assumptions for the Monte Carlo approach equivalent to the assumptions of the classical approach). The frequency distribution of these randomly generated samples provides an estimation of the sampling distribution of the statistic of interest (i.e., $r$ or $F$). For our example, Figure 4 shows the histogram of the values of $r^2$ and $F$ obtained for 1,000 random samples of 6 observations each. The horizontal axes represent the different values of $r^2$ (top panel) and $F$ (bottom panel) obtained for the 1,000 trials, and the vertical axis the number of occurrences of each value of $r^2$ and $F$. For example, the top panel shows that 160 samples (of the 1,000 trials) have a value of $r^2 = .01$, which was between 0 and .01 (this corresponds to the first bar of the histogram in Figure 4).

Figure 4 shows that the number of occurrences of a given value of $r^2$ and $F$ decreases as an inverse function of their magnitude: The greater the value, the less likely it is to obtain it when there is no correlation in the population (i.e., when the null hypothesis is true). However, Figure 4 shows also that the probability of obtaining a large value of $r^2$ or $F$ is not null. In other words, even when the null hypothesis is true, very large values of $r^2$ and $F$ can be obtained.

From now on, this entry focuses on the $F$ distribution, but everything also applies to the $r^2$ distribution. After the sampling distribution has been obtained, the Monte Carlo procedure follows the same steps as the classical approach. Specifically, if the $p$ value for the criterion is smaller than the chosen $\alpha$ level, the null hypothesis can be rejected. Equivalently, a value of $F$ larger than the $\alpha$-level critical value leads one to reject the null hypothesis for this $\alpha$ level.

For our example, we find that 310 random samples (out of 1,000) had a value of $F$ larger than $F = 1.33$, and this corresponds to a probability of $p = .310$ (compare with a value of $p = .313$ for the classical approach). Because this $p$ value is not smaller than $\alpha = .05$, we cannot reject the null hypothesis. Using the critical-value approach leads to the same decision. The empirical critical value for $\alpha = .05$ is equal to 7.5500 (see Figure 4). Because the computed value of $F = 1.33$ is not larger than the 7.5500, we do not reject the null hypothesis.

## Permutation Tests

For both the Monte Carlo and the traditional (i.e., Fisher) approaches, we need to specify the shape of the distribution under the null hypothesis. The Monte Carlo approach can be used with any distribution (but we need to specify which one we want), and the classical approach assumes a normal distribution. An alternative way to look at a null hypothesis test is to evaluate whether the pattern of results for the experiment is a *rare* event by comparing it to all the other patterns of results that could have arisen from these data. This is called a *permutation* test or sometimes a *randomization* test.

This *nonparametric* approach originated with Student and Ronald Fisher, who developed the (now standard) $F$ approach because it was possible then to compute one $F$ but very impractical to compute all the $F$s for all possible permutations. If Fisher could have had access to modern computers, it is likely that permutation tests would be the standard procedure.

So, in order to perform a permutation test, we need to evaluate the probability of finding the value of the statistic of interest (e.g., $r$ or $F$) that we have obtained, compared with all the values we could have obtained by permuting the values of the sample. For our example, we have six observations, and therefore there are

$$6! = 6 \times 5 \times 4 \times 3 \times 2 = 720$$

different possible patterns of results. Each of these patterns corresponds to a given *permutation* of the

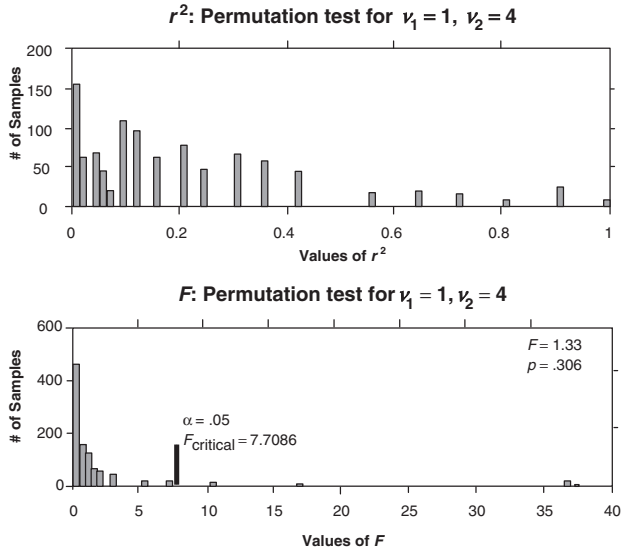



**Figure 5** Histogram of *F* Values Computed From the 6! = 720 Possible Permutations of the Six Scores of the Example

data. For instance, here is a possible permutation of the results for our example:

$$W_1 = 1, W_2 = 3, W_3 = 4, W_4 = 4, W_5 = 5, W_6 = 7$$
$$Y_1 = 8, Y_2 = 10, Y_3 = 16, Y_4 = 12, Y_5 = 10, Y_6 = 4.$$

(Note that we need to permute just *one* of the two series of numbers; here we permuted *Y*). This permutation gives a value of $r_{W.Y} = -.30$ and of $r^2_{WY} = .09$. We computed the value of $r_{W.Y}$ for the remaining 718 permutations. The histogram is plotted in Figure 5, where, for convenience, we have also plotted the histogram of the corresponding *F* values.

For our example, we want to use the permutation test to compute the probability associated with $r^2_{W.Y} = .25$. This is obtained by computing the *proportion* of $r^2_{W.Y}$ larger than .25. We counted 220 $r^2_{W.Y}$ out of 720 larger or equal to .25; this gives a probability of

$$p = \frac{220}{720} = .306.$$

It is interesting to note that this value is very close to the values found with the two other approaches (cf. Fisher distribution $p = .313$ and Monte Carlo $p. = 310$). This similarity is confirmed by comparing Figure 5, where we have plotted the permutation histogram for *F*, with Figure 3, where we have plotted the Fisher distribution.

When the number of observations is small (as is the case for this example with six observations), it is possible to compute all the possible permutations. In this case we have an *exact* permutation test. But the number of permutations grows very fast when the number of observations increases. For example, with 20 observations the total number of permutations is close to $2.4 \times 10^{18}$ (this is a very big number!). Such large numbers obviously prohibit computing all the permutations. Therefore, for samples of large size, we *approximate* the permutation test by using a large number (say 10,000 or 100,000) of random permutations (this approach is sometimes called a *Monte Carlo permutation test*).

## Confidence Intervals

### *Classical Approach*

The value of *r* computed from a sample is an estimation of the correlation of the population from which the sample was obtained. Suppose that we obtain a new sample from the same population and that we compute the value of the coefficient of correlation for this new sample. In what range is this value likely to fall? This question is answered by computing the *confidence interval* of the coefficient of correlation. This gives an upper bound and a lower bound between which the population coefficient of correlation is likely to stand. For example, we want to specify the range of values of $r_{W.Y}$ in which the correlation in the population has a 95% chance of falling.

Using confidence intervals is more general than a null hypothesis test because if the confidence interval excludes the value 0 then we can reject the null hypothesis. But a confidence interval also gives a *range* of probable values for the correlation. Using confidence intervals has another big advantage: We can act *as if* we could accept the null hypothesis. In order to do so, we first compute the confidence interval of the coefficient of correlation and look at the largest magnitude it can have. If we consider that this value is small, then we can say that even if the magnitude of the population correlation is not zero, it is too small to be of interest.

Conversely, we can give more weight to a conclusion if we show that the smallest possible value

for the coefficient of correlation will still be large enough to be impressive.

The problem of computing the confidence interval for $r$ has been explored (once again) by Student and Fisher. Fisher found that the problem was not simple but that it could be simplified by transforming $r$ into another variable called $Z$. This transformation, which is called Fisher's $Z$ transform, creates a new $Z$ variable whose sampling distribution is close to the normal distribution. Therefore, we can use the normal distribution to compute the confidence interval of $Z$, and this will give a lower and a higher bound for the population values of $Z$. Then we can transform these bounds back into $r$ values (using the inverse $Z$ transformation), and this gives a lower and upper bound for the possible values of $r$ in the population.

#### Fisher's Z Transform

Fisher's $Z$ transform is applied to a coefficient of correlation $r$ according to the following formula:

$$Z = \frac{1}{2}[\ln(1 + r) - \ln(1 - r)], \qquad (10)$$

where ln is the *natural* logarithm.

The inverse transformation, which gives $r$ from $Z$, is obtained with the following formula:

$$r = \frac{\exp\{2 \times Z\} - 1}{\exp\{2 \times Z\} + 2}, \qquad (11)$$

where $\exp\{x\}$ means to raise the number $e$ to the power $\{x\}$ (i.e., $\exp\{x\} = e^x$ and $e$ is Euler's constant, which is approximately 2.71828). Most hand calculators can be used to compute both transformations.

Fisher showed that the new $Z$ variable has a sampling distribution that is normal, with a mean of 0 and a variance of $S - 3$. From this distribution we can compute directly the upper and lower bounds of $Z$ and then transform them back into values of $r$.

#### Example

The computation of the confidence interval for the coefficient of correlation is illustrated using the previous example, in which we computed a coefficient of correlation of $r = -.5$ on a sample made of $S = 6$ observations. The procedure can be decomposed into six steps, which are detailed next.

*Step 1.* Before doing any computation, we need to choose an $\alpha$ level that will correspond to the probability of finding the population value of $r$ in the confidence interval. Suppose we chose the value $\alpha = .05$. This means that we want to obtain a confidence interval such that there is a 95% chance, or $(1 - \alpha) = (1 - .05) = .95$, of having the population value being in the confidence interval that we will compute.

*Step 2.* Find in the table of the normal distribution the critical values corresponding to the chosen $\alpha$ level. Call this value $Z_\alpha$. The most frequently used values are

$$Z_{\alpha=.10} = 1.645 \ (\alpha = .10)$$
$$Z_{\alpha=.05} = 1.960 \ (\alpha = .05)$$
$$Z_{\alpha=.01} = 2.575 \ (\alpha = .01)$$
$$Z_{\alpha=001} = 3.325 \ (\alpha = 001).$$

*Step 3.* Transform $r$ into $Z$ using Equation 10. For the present example, with $r = 5$, we fnd that $Z = -0.5493$.

*Step 4.* Compute a quantity called $Q$ as

$$Q = Z_\alpha \times \sqrt{\frac{1}{S - 3}}.$$

For our example we obtain

$$Q = Z_{.05} \times \sqrt{\frac{1}{6 - 3}} = 1.960 \times \sqrt{\frac{1}{3}} = 1.1316.$$

*Step 5.* Compute the lower and upper limits for $Z$ as

$$\text{Lower Limit} = Z_{\text{lower}} = Z - Q$$
$$= -0.5493 - 1.1316 = -1.6809$$

$$\text{Upper Limit} = Z_{\text{upper}} = Z + Q$$
$$= -0.5493 + 1.1316 = 0.5823.$$

*Step 6.* Transform $Z_{\text{lower}}$ and $Z_{\text{upper}}$ into $r_{\text{lower}}$ and $r_{\text{upper}}$. This is done with the use of Equation 11. For the present example, we find that

$$\text{Lower Limit} = r_{\text{lower}} = -.9330$$
$$\text{Upper Limit} = r_{\text{upper}} = .5243.$$

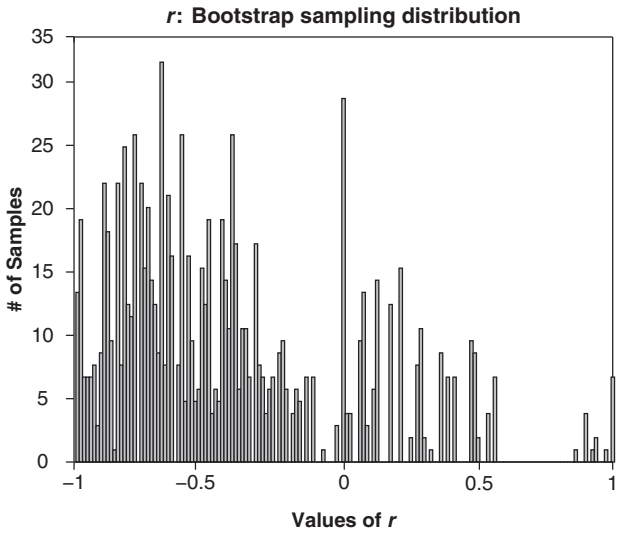



**Figure 6** Histogram of $r_{W.Y}$ Values Computed From 1,000 Bootstrapped Samples Drawn With Replacement From the Data From Our Example

The range of possible values of $r$ is very large: the value of the coefficient of correlation that we have computed could come from a population whose correlation could have been as low as $r_{\text{lower}} = -.9330$ or as high as $r_{\text{upper}} = .5243$. Also, because zero is in the range of possible values, we cannot reject the null hypothesis (which is the conclusion reached with the null hypothesis tests).

It is worth noting that because the $Z$ transformation is nonlinear, the confidence interval is *not* symmetric around $r$.

Finally, current statistical practice recommends the routine use of confidence intervals because this approach is more informative than null hypothesis testing.

## Efron's Bootstrap

A modern Monte Carlo approach for deriving confidence intervals was proposed by Bradley Efron. This approach, called the *bootstrap,* was probably the most important advance for inferential statistics in the second part of the 20th century.

The idea is simple but could be implemented only with modern computers, which explains why it is a recent development. With the bootstrap approach, we treat the sample as if it were the population of interest in order to estimate the sampling distribution of a statistic computed on the sample. Practically this means that in order to estimate the sampling distribution of a statistic, we just need to create bootstrap samples obtained by drawing observations *with replacement* (whereby each observation is put back into the sample after it has been drawn) from the original sample. The distribution of the bootstrap samples is taken as the population distribution. Confidence intervals are then computed from the percentile of this distribution.

For our example, the first bootstrap sample that we obtained comprised the following observations (note that some observations are missing and some are repeated as a consequence of drawing with replacement):

$$
\begin{aligned}
s_1 &= \text{observation } 5,\\
s_2 &= \text{observation } 1,\\
s_3 &= \text{observation } 3,\\
s_4 &= \text{observation } 2,\\
s_4 &= \text{observation } 3,\\
s_6 &= \text{observation } 6.
\end{aligned}
$$

This gives the following values for the first bootstrapped sample obtained by drawing with replacement from our example:

$$
\begin{aligned}
&W_1 = 5, W_2 = 1, W_3 = 4, W_4 = 3, W_5 = 4, W_6 = 7\\
&Y_1 = 8, Y_2 = 16, Y_3 = 12, Y_4 = 10, Y_5 = 12, Y_6 = 10.
\end{aligned}
$$

This bootstrapped sample gives a correlation of $r_{W.Y} = -.73$.

If we repeat the bootstrap procedure for 1,000 samples, we obtain the sampling distribution of $r_{W.Y}$ as shown in Figure 6. From this figure, it is obvious that the value of $r_{W.Y}$ varies a lot with such a small sample (in fact, it covers the whole range of possible values, from $-1$ to $+1$). In order to find the upper and the lower limits of a confidence interval, we look for the corresponding percentiles. For example, if we select a value of $\alpha = .05$, we look at the values of the bootstrapped distribution corresponding to the 2.5th and the 97.5th percentiles. In our example, we find that 2.5% of the values are smaller than $-.9487$ and that 2.5% of the values are larger than .4093. Therefore, these two values constitute the lower

and the upper limits of the 95% confidence interval of the population estimation of $r_{W.Y}$ (cf. the values obtained with Fisher's $Z$ transform of $-.9330$ and $.5243$). Contrary to Fisher's $Z$ transform approach, the bootstrap limits are not dependent on assumptions about the population or its parameters (but it is comforting to see that these two approaches concur for our example). Because the value of 0 is in the confidence interval of $r_{W.Y}$, we cannot reject the null hypothesis. This shows once again that the confidence interval approach provides more information than the null hypothesis approach.

### Shrunken and Adjusted *r*

The coefficient of correlation is a *descriptive* statistic that always overestimates the population correlation. This problem is similar to the problem of the estimation of the variance of a population from a sample. In order to obtain a better estimate of the population, the value $r$ needs to be corrected. The corrected value of $r$ goes under different names: corrected *r,* shrunken *r,* or adjusted *r* (there are some subtle differences between these different appellations, but we will ignore them here) and denote it by $\tilde{r}^2$. Several correction formulas are available. The one most often used estimates the value of the population correlation as

$$\tilde{r}^2 = 1 - \left[(1 - r^2)\left(\frac{S-1}{S-2}\right)\right]. \qquad (12)$$

For our example, this gives

$$\tilde{r}^2 = 1 - \left[(1 - r^2)\left(\frac{S-1}{S-2}\right)\right] = 1 - \left[(1 - .25) \times \frac{5}{4}\right]$$
$$= 1 - \left[.75 \times \frac{5}{4}\right] = 0.06.$$

With this formula, we find that the estimation of the population correlation drops from $r = -.50$ to $\tilde{r}^2 = -\sqrt{\tilde{r}^2} = -\sqrt{.06} = -.24$.

### Particular Cases of the Coefficient of Correlation

Mostly for historical reasons, some specific cases of the coefficient of correlation have their own names (in part because these special cases lead to simplified computational formulas). Specifically, when both variables are ranks (or transformed into ranks), we obtain the *Spearman rank correlation coefficient* (a related transformation will provide the *Kendall rank correlation coefficient*); when both variables are dichotomous (i.e., they take only the values 0 and 1), we obtain the *phi coefficient of correlation;* and when only one of the two variables is dichotomous, we obtain the *point-biserial coefficient*.

*Hervé Abdi and Lynne J. Williams*

***See also*** Coefficient of Concordance; Confidence Intervals

**Further Readings**

Abdi, H., Edelman, B., Valentin, D., & Dowling, W. J. (2009). *Experimental design and analysis for psychology.* Oxford, UK: Oxford University Press.

Cohen, J., & Cohen, P. (1983) *Applied multiple regression/correlation analysis for the social sciences.* Hillsdale, NJ: Lawrence Erlbaum.

Darlington, R. B. (1990). *Regression and linear models.* New York: McGraw-Hill.

Edwards, A. L. (1985). *An introduction to linear regression and correlation.* New York: Freeman.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research.* New York: Harcourt Brace.

## Cohen's *d* Statistic

Cohen's *d* statistic is a type of *effect size*. An effect size is a specific numerical nonzero value used to represent the extent to which a null hypothesis is false. As an effect size, Cohen's *d* is typically used to represent the magnitude of differences between two (or more) groups on a given variable, with larger values representing a greater differentiation between the two groups on that variable. When comparing means in a scientific study, the reporting of an effect size such as Cohen's *d* is considered complementary to the reporting of results from a test of statistical significance. Whereas the test of statistical significance is used to suggest whether a null hypothesis is true (no difference exists between Populations A and B for a specific phenomenon) or false (a difference exists between

Populations A and B for a specific phenomenon), the calculation of an effect size estimate is used to represent the degree of difference between the two populations in those instances for which the null hypothesis was deemed false. In cases for which the null hypothesis is false (i.e., rejected), the results of a test of statistical significance imply that reliable differences exist between two populations on the phenomenon of interest, but test outcomes do not provide any value regarding the extent of that difference. The calculation of Cohen's *d* and its interpretation provide a way to estimate the actual size of observed differences between two groups, namely, whether the differences are small, medium, or large.

## Calculation of Cohen's *d* Statistic

Cohen's *d* statistic is typically used to estimate between-subjects effects for grouped data, consistent with an analysis of variance framework. Often, it is employed within experimental contexts to estimate the differential impact of the experimental manipulation across conditions on the dependent variable of interest. The dependent variable must represent continuous data; other effect size measures (e.g., Pearson family of correlation coefficients, odds ratios) are appropriate for noncontinuous data.

### *General Formulas*

Cohen's *d* statistic represents the standardized mean differences between groups. Similar to other means of standardization such as *z* scoring, the effect size is expressed in standard score units. In general, Cohen's *d* is defined as

$$d = \frac{\mu_1 - \mu_2}{\sigma_\varepsilon}, \tag{1}$$

where $d$ represents the effect size, $\mu_1$ and $\mu_2$ represent the two population means, and $\sigma_\varepsilon$ represents the pooled within-group population standard deviation. In practice, these population parameters are typically unknown and estimated by means of sample statistics:

$$\hat{d} = \frac{\overline{Y}_1 - \overline{Y}_2}{S_p}. \tag{2}$$

The population means are replaced with sample means ($\overline{Y}_j$), and the population standard deviation is replaced with $S_p$, the pooled standard deviation from the sample. The pooled standard deviation is derived by weighing the variance around each sample mean by the respective sample size.

### *Calculation of the Pooled Standard Deviation*

Although computation of the difference in sample means is straightforward in Equation 2, the pooled standard deviation may be calculated in a number of ways. Consistent with the traditional definition of a standard deviation, this statistic may be computed as

$$S_p = \sqrt{\frac{\sum (n_j - 1)s_j^2}{\sum (n_j)}}, \tag{3}$$

where $n_j$ represents the sample sizes for $j$ groups and $s_j^2$ represents the variance (i.e., squared standard deviation) of the $j$ samples. Often, however, the pooled sample standard deviation is corrected for bias in its estimation of the corresponding population parameter, $\sigma_\varepsilon$. Equation 4 denotes this correction of bias in the sample statistic (with the resulting effect size often referred to as Hedge's *g*):

$$S_p = \sqrt{\frac{\sum (n_j - 1)s_j^2}{\sum (n_j - 1)}}. \tag{4}$$

When simply computing the pooled standard deviation across two groups, this formula may be reexpressed in a more common format. This formula is suitable for data analyzed with a two-way analysis of variance, such as a treatment–control contrast:

$$\begin{aligned} S_p &= \sqrt{\frac{(n_1 - 1)(s_1^2) + (n_2 - 1)(s_2^2)}{(n_1 - 1) + (n_2 - 1)}} \\ &= \sqrt{\frac{(n_1 - 1)(s_1^2) + (n_2 - 1)(s_2^2)}{(n_1 + n_2 - 2)}}. \end{aligned} \tag{5}$$

The formula may be further reduced to the average of the sample variances when sample sizes are equal:

$$S_p = \sqrt{\frac{s_j^2}{j}}. \quad (6)$$

or

$$S_p = \sqrt{\frac{s_1^2 + s_2^2}{2}}. \quad (7)$$

in the case of two groups.

Other means of specifying the denominator for Equation 2 are varied. Some formulas use the average standard deviation across groups. This procedure disregards differences in sample size in cases of unequal *n* when one is weighing sample variances and may or may not correct for sample bias in estimation of the population standard deviation. Further formulas employ the standard deviation of the control or comparison condition (an effect size referred to as Glass's $\Delta$). This method is particularly suited when the introduction of treatment or other experimental manipulation leads to large changes in group variance. Finally, more complex formulas are appropriate when calculating Cohen's *d* from data involving cluster randomized or nested research designs. The complication partially arises because of the three available variance statistics from which the pooled standard deviation may be computed: the within-cluster variance, the between-cluster variance, or the total variance (combined between- and within-cluster variance). Researchers must select the variance statistic appropriate for the inferences they wish to draw.

### *Expansion Beyond Two-Group Comparisons: Contrasts and Repeated Measures*

Cohen's *d* always reflects the standardized difference between two means. The means, however, are not restricted to comparisons of two independent groups. Cohen's *d* may also be calculated in multigroup designs when a specific contrast is of interest. For example, the average effect across two alternative treatments may be compared with a control. The value of the contrast becomes the numerator as specified in Equation 2, and the pooled standard deviation is expanded to include all *j* groups specified in the contrast (Equation 4).

A similar extension of Equations 2 and 4 may be applied to repeated measures analyses. The difference between two repeated measures is divided by the pooled standard deviation across the *j* repeated measures. The same formula may also be applied to simple contrasts within repeated measures designs, as well as interaction contrasts in mixed (between- and within-subjects factors) or split-plot designs. Note, however, that the simple application of the pooled standard deviation formula does not take into account the correlation between repeated measures. Researchers disagree as to whether these correlations ought to contribute to effect size computation; one method of determining Cohen's *d* while accounting for the correlated nature of repeated measures involves computing *d* from a paired *t* test.

### *Additional Means of Calculation*

Beyond the formulas presented above, Cohen's *d* may be derived from other statistics, including the Pearson family of correlation coefficients (*r*), *t* tests, and *F* tests. Derivations from *r* are particularly useful, allowing for translation among various effect size indices. Derivations from other statistics are often necessary when raw data to compute Cohen's *d* are unavailable, such as when conducting a meta-analysis of published data. When *d* is derived as in Equation 3, the following formulas apply:

$$d = \frac{2r}{\sqrt{1 - r^2}}, \quad (8)$$

$$d = \frac{t(n_1 + n_2)}{\sqrt{df}\sqrt{n_1 n_2}}, \quad (9)$$

and

$$d = \frac{\sqrt{F}(n_1 + n_2)}{\sqrt{df}\sqrt{n_1 n_2}}. \quad (10)$$

Note that Equation 10 applies only for *F* tests with 1 degree of freedom (*df*) in the numerator; further formulas apply when $df > 1$.

When *d* is derived as in Equation 4, the following formulas ought to be used:

$$d = \frac{2r}{\sqrt{1 - r^2}}\sqrt{\frac{df(n_1 + n_2)}{n_1 n_2}} \quad (11)$$

$$d = t\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{12}$$

$$d = \sqrt{F\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}. \tag{13}$$

Again, Equation 13 applies only to instances in which the numerator $df = 1$.

These formulas must be corrected for the correlation ($r$) between dependent variables in repeated measures designs. For example, Equation 12 is corrected as follows:

$$d = t\sqrt{\left(\frac{(1-r)}{n_1} + \frac{(1-r)}{n_2}\right)}. \tag{14}$$

Finally, conversions between effect sizes computed with Equations 3 and 4 may be easily accomplished:

$$d_{eq3} = d_{eq4}\sqrt{\frac{n_1 + n_2}{(n_1 + n_2 - 2)}} \tag{15}$$

and

$$d_{eq4} = \frac{d_{eq3}}{\sqrt{\frac{n_1+n_2}{(n_1+n_2-2)}}}. \tag{16}$$

### *Variance and Confidence Intervals*

The estimated variance of Cohen's $d$ depends on how the statistic was originally computed. When sample bias in the estimation of the population pooled standard deviation remains uncorrected (Equation 3), the variance is computed in the following manner:

$$s_d = \left(\frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2 - 2)}\right)\left(\frac{n_1 + n_2}{n_1 + n_2 - 2}\right). \tag{17}$$

A simplified formula is employed when sample bias is corrected as in Equation 4:

$$s_d = \frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2 - 2)}. \tag{18}$$

Once calculated, the effect size variance may be used to compute a confidence interval (CI) for the statistic to determine statistical significance:

$$\text{CI} = d \pm z(s_d). \tag{19}$$

The $z$ in the formula corresponds to the $z$-score value on the normal distribution corresponding to the desired probability level (e.g., 1.96 for a 95% CI). Variances and CIs may also be obtained through bootstrapping methods.

## Interpretation

Cohen's $d$, as a measure of effect size, describes the overlap in the distributions of the compared samples on the dependent variable of interest. If the two distributions overlap completely, one would expect no mean difference between them (i.e., $\overline{Y}_1 - \overline{Y}_2 = 0$). To the extent that the distributions do not overlap, the difference ought to be greater than zero (assuming $\overline{Y}_1 > \overline{Y}_2$).

Cohen's $d$ may be interpreted in terms of both statistical significance and magnitude, with the latter the more common interpretation. Effect sizes are statistically significant when the computed CI does not contain zero. This implies less than perfect overlap between the distributions of the two groups compared. Moreover, the significance testing implies that this difference from zero is reliable, or not due to chance (excepting Type I errors). While significance testing of effect sizes is often undertaken, however, interpretation based solely on statistical significance is not recommended. Statistical significance is reliant not only on the size of the effect but also on the size of the sample. Thus, even large effects may be deemed unreliable when insufficient sample sizes are utilized.

Interpretation of Cohen's $d$ based on the magnitude is more common than interpretation based on statistical significance of the result. The magnitude of Cohen's $d$ indicates the extent of nonoverlap between two distributions, or the disparity of the mean difference from zero. Larger numeric values of Cohen's $d$ indicate larger effects or greater differences between the two means. Values may be positive or negative, although the sign merely indicates whether the first or second mean in the numerator was of greater magnitude (see Equation 2). Typically, researchers choose to subtract the smaller mean from the larger, resulting in a positive

effect size. As a standardized measure of effect, the numeric value of Cohen's *d* is interpreted in standard deviation units. Thus, an effect size of $d = 0.5$ indicates that two group means are separated by one-half standard deviation or that one group shows a one-half standard deviation advantage over the other.

The magnitude of effect sizes is often described nominally as well as numerically. Jacob Cohen defined effects as small ($d = 0.2$), medium ($d = 0.5$), or large ($d = 0.8$). These rules of thumb were derived after surveying the behavioral sciences literature, which included studies in various disciplines involving diverse populations, interventions or content under study, and research designs. Cohen, in proposing these benchmarks in a 1988 text, explicitly noted that they are arbitrary and thus ought not be viewed as absolute. However, as occurred with use of .05 as an absolute criterion for establishing statistical significance, Cohen's benchmarks are oftentimes interpreted as absolutes, and as a result, they have been criticized in recent years as outdated, atheoretical, and inherently nonmeaningful. These criticisms are especially prevalent in applied fields in which medium-to-large effects prove difficult to obtain and smaller effects are often of great importance. The small effect of $d = 0.07$, for instance, was sufficient for physicians to begin recommending aspirin as an effective method of preventing heart attacks. Similar small effects are often celebrated in intervention and educational research, in which effect sizes of $d = 0.3$ to $d = 0.4$ are the norm. In these fields, the practical importance of reliable effects is often weighed more heavily than simple magnitude, as may be the case when adoption of a relatively simple educational approach (e.g., discussing vs. not discussing novel vocabulary words when reading storybooks to children) results in effect sizes of $d = 0.25$ (consistent with increases of one-fourth of a standard deviation unit on a standardized measure of vocabulary knowledge).

Critics of Cohen's benchmarks assert that such practical or substantive significance is an important consideration beyond the magnitude and statistical significance of effects. Interpretation of effect sizes requires an understanding of the context in which the effects are derived, including the particular manipulation, population, and dependent measure(s) under study. Various alternatives to Cohen's rules of thumb have been proposed. These include comparisons with effects sizes based on (a) normative data concerning the typical growth, change, or differences between groups prior to experimental manipulation; (b) those obtained in similar studies and available in the previous literature; (c) the gain necessary to attain an a priori criterion; and (d) cost–benefit analyses.

## Cohen's *d* in Meta-Analyses

Cohen's *d,* as a measure of effect size, is often used in individual studies to report and interpret the magnitude of between-group differences. It is also a common tool used in meta-analyses to aggregate effects across different studies, particularly in meta-analyses involving study of between-group differences, such as treatment studies. A meta-analysis is a statistical synthesis of results from independent research studies (selected for inclusion based on a set of predefined commonalities), and the unit of analysis in the meta-analysis is the data used for the independent hypothesis test, including sample means and standard deviations, extracted from each of the independent studies. The statistical analyses used in the meta-analysis typically involve (a) calculating the Cohen's *d* effect size (standardized mean difference) on data available within each independent study on the target variable(s) of interest and (b) combining these individual summary values to create pooled estimates by means of any one of a variety of approaches (e.g., Rebecca DerSimonian and Nan Laird's random effects model, which takes into account variations among studies on certain parameters). Therefore, the methods of the meta-analysis may rely on use of Cohen's *d* as a way to extract and combine data from individual studies. In such meta-analyses, the reporting of results involves providing average *d* values (and CIs) as aggregated across studies.

In meta-analyses of treatment outcomes in the social and behavioral sciences, for instance, effect estimates may compare outcomes attributable to a given treatment (Treatment X) as extracted from and pooled across multiple studies in relation to an alternative treatment (Treatment Y) for Outcome Z using Cohen's *d* (e.g., $d = 0.21$, CI = 0.06, 1.03). It is important to note that the meaningfulness of this result, in that Treatment X is, on average, associated with an improvement of about

one-fifth of a standard deviation unit for Outcome Z relative to Treatment Y, must be interpreted in reference to many factors to determine the actual significance of this outcome. Researchers must, at the least, consider whether the one-fifth of a standard deviation unit improvement in the outcome attributable to Treatment X has any practical significance.

*Shayne B. Piasta and Laura M. Justice*

***See also*** Analysis of Variance (ANOVA); Effect Size, Measures of; Mean Comparisons; Meta-Analysis; *Statistical Power Analysis for the Behavioral Sciences*

**Further Readings**

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Cooper, H., & Hedges, L. V. (1994). *The handbook of research synthesis*. New York: Russell Sage Foundation.

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational & Behavioral Statistics, 32,* 341–370.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2007, July). *Empirical benchmarks for interpreting effect sizes in research*. New York: MDRC.

Ray, J. W., & Shadish, W. R. (1996). How interchangeable are different estimators of effect size? *Journal of Consulting & Clinical Psychology, 64,* 1316–1325.

Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.

# Cohen's *f* Statistic

Effect size is a measure of the strength of the relationship between variables. Cohen's *f* statistic is one appropriate effect size index to use for a one-way analysis of variance (ANOVA). Cohen's *f* is a measure of a kind of standardized average effect in the population across all the levels of the independent variable.

Cohen's *f* can take on values between zero, when the population means are all equal, and an indefinitely large number as standard deviation of means increases relative to the average standard deviation within each group. Jacob Cohen has suggested that the values of 0.10, 0.25, and 0.40 represent small, medium, and large effect sizes, respectively.

## Calculation

Cohen's *f* is calculated as

$$f = \sigma_m/\sigma, \tag{1}$$

where $\sigma_m$ is the standard deviation (*SD*) of population means ($m_i$) represented by the samples and $\sigma$ is the common within-population *SD*; $\sigma = MSE^{1/2}$. *MSE* is the mean square of error (within groups) from the overall ANOVA *F* test. It is based on the deviation of the population means from the mean of the combined populations or the mean of the means (*M*).

$$\sigma_m = \left[\sum (m_i - M)^2/k\right]^{1/2} \tag{2}$$

for equal sample sizes and

$$\sigma_m = \left[\sum n_i(m_i - M)^2/N\right]^{1/2} \tag{3}$$

for unequal sample sizes.

## Examples

### *Example 1*

Table 1 provides descriptive statistics for a study with four groups and equal sample sizes. ANOVA results are shown. The calculations below result in an estimated *f* effect size of .53, which is considered large by Cohen standards. An appropriate interpretation is that about 50% of the variance in the dependent variable (physical health) is explained by the independent variable (presence or absence of mental or physical illnesses at age 16).

$$\sigma_m = \left[\sum (m_i - M)^2/k\right]^{1/2} = [((71.88 - 62.74)^2 + (66.08 - 62.74)^2 + (58.44 - 62.74)^2 + 54.58 - 62.74)^2)/4]^{1/2} = 6.70$$

$$f = \sigma_m/\sigma = 6.70/161.29^{1/2} = 6.70/12.7 = 0.53$$

**Table 1** Association of Mental Disorders and Physical Illnesses at a Mean Age of 16 Years With Physical Health at a Mean Age of 33 Years: Equal Sample Sizes

| *Group* | *n* | *M* | *SD* |
|---|---|---|---|
| Reference group (no disorders) | 80 | 71.88 | 5.78 |
| Mental disorder only | 80 | 66.08 | 19.00 |
| Physical illness only | 80 | 58.44 | 8.21 |
| Physical illness and mental disorder | 80 | 54.58 | 13.54 |
| Total | 320 | 62.74 | 14.31 |

*Analysis of Variance*

| *Source* | *Sum of Squares* | *df* | *Mean Square* | *F* | *p* |
|---|---|---|---|---|---|
| Between groups | 14,379.93 | 3 | 4,793.31 | 29.72 | 0.00 |
| Within groups | 50,967.54 | 316 | 161.29 | | |
| Total | 65,347.47 | 319 | | | |

*Source:* Adapted from Chen, H., Cohen, P., Kasen, S., Johnson, J. G., Berenson, K., & Gordon, K. (2006). Impact of adolescent mental disorders and physical illnesses on quality of life 17 years later. *Archives of Pediatrics & Adolescent Medicine, 160,* 93–99.

**Table 2** Association of Mental Disorders and Physical Illnesses at a Mean Age of 16 Years With Physical Health at a Mean Age of 33 Years: Unequal Sample Sizes

| *Group* | *n* | *M* | *SD* |
|---|---|---|---|
| Reference group (no disorders) | 256 | 72.25 | 17.13 |
| Mental disorder only | 89 | 68.16 | 21.19 |
| Physical illness only | 167 | 66.68 | 18.58 |
| Physical illness and mental disorder | 96 | 57.67 | 18.86 |
| Total | 608 | 67.82 | 19.06 |

*Analysis of Variance*

| *Source* | *Sum of Squares* | *df* | *Mean Square* | *F* | *p* |
|---|---|---|---|---|---|
| Between groups | 15,140.20 | 3 | 5,046.73 | 14.84 | 0.00 |
| Within groups | 205,445.17 | 604 | 340.14 | | |
| Total | 220,585.37 | 607 | | | |

*Source:* Adapted from Chen, H., Cohen, P., Kasen, S., Johnson, J. G., Berenson, K., & Gordon, K. (2006). Impact of adolescent mental disorders and physical illnesses on quality of life 17 years later. *Archives of Pediatrics & Adolescent Medicine, 160,* 93–99.

*Note:* Adapted from total sample size of 608 by choosing 80 subjects for each group.

### *Example 2*

Table 2 provides descriptive statistics for a study very similar to the one described in Example 1. There are four groups, but the sample sizes are unequal. ANOVA results are shown. The calculations for these samples result in an estimated *f* effect size of .27, which is considered medium by Cohen standards. For this study, an appropriate interpretation is that about 25% of the variance in the dependent variable (physical health) is explained by the independent variable (presence or absence of mental or physical illnesses at age 16).

$$\sigma_m = [\Sigma_{n_i}(m_i - M)^2/N]^{1/2} = [(256 * (72.25 - 67.82)^2 + 89 * (68.16 - 67.82)^2 + 167 * (66.68 - 67.82)^2 + 96 * (57.67 - 67.82)^2)/608]^{1/2} = 4.99$$

$$f = \sigma_m/\sigma = 4.99/340.14^{1/2} = 4.99/18.44 = 0.27$$

### Cohen's *f* and *d*

Cohen's *f* is an extension of Cohen's *d*, which is the appropriate measure of effect size to use for a *t* test. Cohen's *d* is the difference between two group means divided by the pooled *SD* for the two groups. The relationship between *f* and *d* when one is comparing two means (equal sample sizes) is $d = 2f$. If Cohen's $f = 0.1$, the *SD* of $k (k \geq 2)$ population means is one tenth as large as the *SD* of the observations within the populations. For $k =$ two populations, this effect size indicates a small difference between the two populations: $d = 2f = 2 * 0.10 = 0.2$.

Cohen's *f* in Equation 1 is positively biased because the sample means in Equation 2 or 3 are likely to vary more than do the population means. One can use the following equation from Scott Maxwell and Harold Delaney to calculate an adjusted Cohen's *f:*

$$f_{\text{adj}} = [(k-1)(F-1)/N]^{1/2} \qquad (4)$$

Applying Equation 4 to the data in Table 1 yields

$$f_{adj} = [(k-1)(F-1)/N]^{1/2}$$
$$= [(4-1)(29.72-1)/320]^{1/2} = 0.52.$$

For Table 2,

$$f_{adj} = [(k-1)(F-1)/N]^{1/2}$$
$$= [(4-1)(14.84-1)/608]^{1/2} = 0.26.$$

*Sophie Chen and Henian Chen*

***See also*** Cohen's *d* Statistic; Effect Size, Measures of

#### Further Readings

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

## Cohen's Kappa

Cohen's Kappa coefficient ($\kappa$) is a statistical measure of the degree of agreement or concordance between two independent raters that takes into account the possibility that agreement could occur by chance alone.

Like other measures of interrater agreement, $\kappa$ is used to assess the reliability of different raters or measurement methods by quantifying their consistency in placing individuals or items in two or more mutually exclusive categories. For instance, in a study of developmental delay, two pediatricians may independently assess a group of toddlers and classify them with respect to their language development into either "delayed for age" or "not delayed." One important aspect of the utility of this classification is the presence of good agreement between the two raters. Agreement between two raters could be simply estimated as the percentage of cases in which both raters agreed. However, a certain degree of agreement is expected by chance alone. In other words, two raters could still agree on some occasions even if they were randomly assigning individuals into either category.

In situations in which there are two raters and the categories used in the classification system have no natural order (e.g., delayed vs. not delayed; present vs. absent), Cohen's $\kappa$ can be used to quantify the degree of agreement in the assignment of these categories beyond what would be expected by random guessing or chance alone.

### Calculation

Specifically, $\kappa$ can be calculated using the following equation:

$$\kappa = \frac{p_o - p_e}{1 - p_e},$$

where $P_o$ is the proportion of the observed agreement between the two raters, and $P_e$ is the proportion of rater agreement expected by chance alone. A $\kappa$ of $+1$ indicates complete agreement, whereas a $\kappa$ of 0 indicates that there is no agreement between the raters beyond that expected by random guessing or chance alone. A negative $\kappa$ indicates that the agreement was less than expected by

**Table 1** Data From the Hypothetical Study Described in the Text: Results of Assessments of Developmental Delay Made by Two Pediatricians

| | *Rater 2* | | |
|---|---|---|---|
| *Rater 1* | *Delayed* | *Not Delayed* | *Total* |
| Delayed | 20 | 3 | 23 |
| Not delayed | 7 | 70 | 77 |
| Total | 27 | 73 | 100 |

**Table 2** Landis and Koch Interpretation of Cohen's $\kappa$

| *Cohen's Kappa* | *Degree of Agreement* |
|---|---|
| $< 0.20$ | Poor |
| 0.21–0.40 | Fair |
| 0.41–0.60 | Moderate |
| 0.61–0.80 | Good |
| 0.81–1.00 | Very good |

*Source:* Landis & Koch, 1977.

chance, with $\kappa$ of $-1.0$ indicating perfect disagreement beyond what would be expected by chance.

To illustrate the use of the equation, let us assume that the results of the assessments made by the two pediatricians in the above-mentioned example are as shown in Table 1. The two raters agreed on the classification of 90 toddlers (i.e., $P_o$ is 0.90). To calculate the probability of the expected agreement ($P_e$), we first calculate the probability that both raters would have classified a toddler as delayed if they were merely randomly classifying toddlers to this category. This could be obtained by multiplying the marginal probabilities of the delayed category, that is, $(23 \div 100) \times (27 \div 100) = 0.062$. Similarly, the probability that both raters would have randomly classified a toddler as not delayed is $(77 \div 100) \times (73 \div 100) = 0.562$. Therefore, the total agreement expected by chance alone ($P_e$) is $0.562 + 0.062 = 0.624$. Using the equation, $\kappa$ is equal to 0.73.

Richard Landis and Gary Koch have proposed the following interpretation for estimates of $\kappa$ (Table 2). Although arbitrary, this classification is widely used in the medical literature. According to this classification, the agreement between the two pediatricians in the above example is "good."

It should be noted that $\kappa$ is a summary measure of the agreement between two raters and cannot therefore be used to answer all the possible questions that may arise in a reliability study. For instance, it might be of interest to determine whether disagreement between the two pediatricians in the above example was more likely to occur when diagnosing developmental delay than when diagnosing normal development or vice versa. However, $\kappa$ cannot be used to address this question, and alternative measures of agreement are needed for that purpose.

## Limitations

It has been shown that the value of $\kappa$ is sensitive to the prevalence of the trait or condition under investigation (e.g., developmental delay in the above example) in the study population. Although two raters may have the same degree of agreement, estimates of $\kappa$ might be different for a population in which the trait under study is very common compared with another population in which the trait is less prevalent.

Cohen's $\kappa$ does not take into account the seriousness of the disagreement. For instance, if a trait is rated on a scale that ranges from 1 to 5, all disagreements are treated equally, whether the raters disagreed by 1 point (e.g., 4 vs. 5) or by 4 points on the scale (e.g., 1 vs. 5). Cohen introduced *weighted* $\kappa$ for use with such ordered scales. By assigning weights to each disagreement pair, it is possible to incorporate a measure of the seriousness of the disagreement, for instance by assigning larger weights to more severe disagreement.

Both $\kappa$ and weighted $\kappa$ are limited to cases in which there are only two raters. Alternative measures of agreements (e.g., Fleiss $\kappa$) can be used in situations in which there are more than two raters and in which study subjects are not necessarily always rated by the same pair of raters.

*Salaheddin M. Mahmud*

*See also* Interrater Reliability; Reliability

**Further Readings**

Banerjee, M., Capozzoli, M., McSweeney, L., & Sinha, D. (1999). Beyond kappa: A review of interrater agreement measures. *Canadian Journal of Statistics, 27,* 3–23.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational & Psychological Measurement, 1,* 37–46.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed.). New York: Wiley.

Gwet, K. (2002). Inter-rater reliability: Dependency on trait prevalence and marginal homogeneity. *Statistical Methods for Inter-Rater Reliability Assessment Series, 2,* 1–9.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159–174.

## Cohort Design

In epidemiology, a cohort design or cohort study is a nonexperimental study design that involves comparing the occurrence of a disease or condition in two or more groups (or *cohorts*) of people that differ on a certain characteristic, risk factor, or exposure. The disease, state, or condition under study is often referred to as the *outcome,* whereas the characteristic, risk factor, or exposure is often referred to as the *exposure.* A cohort study is one of two principal types of nonexperimental study designs used to study the causes of disease. The other is the case–control design, in which cases of the disease under study are compared with respect to their past exposure with a similar group of individuals who do not have the disease.

*Cohort* (from the Latin *cohors,* originally a unit of a Roman legion) is the term used in epidemiology to refer to a group of individuals who share a common characteristic; for example, they may all belong to the same ethnic or age group or be exposed to the same risk factor (e.g., radiation or soil pollution).

The cohort study is a relatively recent innovation. The first cohort studies were used to confirm the link between smoking and lung cancer that had been observed initially in earlier case–control studies. Cohort studies also formed the basis for much of the early progress in understanding occupational diseases. Cohort studies based on data derived from company records and vital records led to the identification of many environmental and occupational risk factors. Several major cohort studies with follow-up that spanned decades have made significant contributions to our understanding of the causes of several common chronic diseases. Examples include the Framingham Heart Study, the Tecumseh Community Health Study, the British Doctors Study, and the Nurses' Health Study.

In a classic cohort study, individuals who are initially free of the disease being researched are enrolled into the study, and individuals are each categorized into one of two groups according to whether they have been exposed to the suspected risk factor. One group, called the *exposed group,* includes individuals known to have the characteristic or risk factor under study. For instance, in a cohort study of the effect of smoking on lung cancer, the exposed group may consist of known smokers. The second group, the *unexposed group,* will comprise a comparable group of individuals who are also free of the disease initially but are nonsmokers. Both groups are then followed up for a predetermined period of time or until the occurrence of disease or death. Cases of the disease (lung cancer in this instance) occurring among both groups are identified in the same way for both groups. The number of people diagnosed with the disease in the exposed group is compared with that among the unexposed group to estimate the relative risk of disease due to the exposure or risk factor. This type of design is sometimes called a *prospective cohort study.*

In a *retrospective (historical) cohort study,* the researchers use existing records or electronic databases to identify individuals who were exposed at a certain point in the past and then "follow" them up to the present. For instance, to study the effect of exposure to radiation on cancer occurrence among workers in a uranium mine, the researcher may use employee radiation exposure records to categorize workers into those who were exposed to radiation and those who were not at a certain date in the past (e.g., 10 years ago). The medical records of each employee are then searched to identify those employees who were diagnosed with

cancer from that date onward. Like prospective cohort designs, the frequency of occurrence of the disease in the exposed group is compared with that within the unexposed group in order to estimate the relative risk of disease due to radiation exposure. When accurate and comprehensive records are available, this approach could save both time and money. But unlike the classic cohort design, in which information is collected prospectively, the researcher employing a retrospective cohort design has little control over the quality and availability of information.

Cohort studies could also be classified as closed or open cohort studies. In a *closed cohort study,* cohort membership is decided at the onset of the study, and no additional participants are allowed to join the cohort once the study starts. For example, in the landmark British Doctors Study, participants were male doctors who were registered for medical practice in the United Kingdom in 1951. This cohort was followed up with periodic surveys until 2001. This study provided strong evidence for the link between smoking and several chronic diseases, including lung cancer.

In an *open* (*dynamic*) *cohort study,* the cohort membership may change over time as additional participants are permitted to join the cohort, and they are followed up in a fashion similar to that of the original participants. For instance, in a prospective study of the effects of radiation on cancer occurrence among uranium miners, newly recruited miners are enrolled in the study cohort and are followed up in the same way as those miners who were enrolled at the inception of the cohort study.

Regardless of their type, cohort studies are distinguished from other epidemiological study designs by having all the following features:

- The study group or groups are observed over time for the occurrence of the study outcome.
- The study group or groups are defined on the basis of whether they have the exposure at the start or during the observation period before the occurrence of the outcome. Therefore, in a cohort study, it is always clear that the exposure has occurred before the outcome.
- Cohort studies are observational or nonexperimental studies. Unlike clinical trials, the cohort study design does not usually involve manipulating the exposure under study in any way that changes the exposure status of the participants.

## Advantages and Disadvantages

Cohort studies are used instead of experimental study designs, such as clinical trials, when experiments are not feasible for practical or ethical reasons, such as when investigating the effects of a potential cause of disease.

In contrast to case–control studies, the design of cohort studies is intuitive, and their results are easier to understand by nonspecialists. Furthermore, the temporal sequence of events in a cohort study is clear because it is always known that the exposure has occurred before the disease. In case–control and cross-sectional studies, it is often unclear whether the suspected exposure has led to the disease or the other way around.

In prospective cohort studies, the investigator has more control over what information is to be collected and at what intervals. As a result, the prospective cohort design is well suited for studying chronic diseases because it permits fuller understanding of the disease's natural history.

In addition, cohort studies are typically better than case–control studies in studying rare exposures. For instance, case–control designs are not practical in studying occupational exposures that are rare in the general population, as when exposure is limited to a small cohort of workers in a particular industry. Another advantage of cohort studies is that multiple diseases and conditions related to the same exposure could be easily examined in one study.

On the other hand, cohort studies tend to be more expensive and take longer to complete than other nonexperimental designs. Generally, case–control and retrospective cohort studies are more efficient and less expensive than the prospective cohort studies. The prospective cohort design is not suited for the study of rare diseases, because prospective cohort studies require following up a large number of individuals for a long time. Maintaining participation of study subjects over time is a challenge, and selective dropout from the study (or loss to follow-up) may result in biased results. Because of lack of randomization, cohort

studies are more potentially subject to bias and confounding than experimental studies are.

## Design and Implementation

The specifics of cohort study design and implementation depend on the aim of the study and the nature of the risk factors and diseases under study. However, most prospective cohort studies begin by assembling one or more groups of individuals. Often members of each group share a well-defined characteristic or exposure. For instance, a cohort study of the health effects of uranium exposure may begin by recruiting all uranium miners employed by the same company, whereas a cohort study of the health effects of exposure to soil pollutants from a landfill may include all people living within a certain distance from the landfill. Several major cohort studies have recruited all people born in the same year in a city or province (birth cohorts). Others have included all members of a professional group (e.g., physicians or nurses), regardless of where they lived or worked. Yet others were based on a random sample of the population. Cohort studies of the natural history of disease may include all people diagnosed with a precursor or an early form of the disease and then followed up as their disease progressed.

The next step is to gather information on the exposure under investigation. Cohort studies can be used to examine exposure to external agents, such as radiation, second-hand smoke, an infectious agent, or a toxin. But they can also be used to study the health effects of internal states (e.g., possession of a certain gene), habits (e.g., smoking or physical inactivity), or other characteristics (e.g., level of income or educational status). The choice of the appropriate exposure measurement method is an important design decision and depends on many factors, including the accuracy and reliability of the available measurement methods, the feasibility of using these methods to measure the exposure for all study participants, and the cost.

In most prospective cohort studies, baseline information is collected on all participants as they join the cohort, typically using self-administered questionnaires or phone or in-person interviews. The nature of the collected information depends on the aim of the study but often includes detailed information on the exposure(s) under investigation. In addition, information on demographic and socioeconomic factors (e.g., age, gender, and occupation) is often collected. As in all observational studies, information on potential confounders, factors associated with both the exposure and outcome under study that could confuse the interpretation of the results, is also collected. Depending on the type of the exposure under study, the study design may also include medical examinations of study participants, which may include clinical assessment (e.g., measuring blood pressure), laboratory testing (e.g., measuring blood sugar levels or testing for evidence for an infection with a certain infectious agent), or radiological examinations (e.g., chest x-rays). In some studies, biological specimens (e.g., blood or serum specimens) are collected and stored for future testing.

Follow-up procedures and intervals are also important design considerations. The primary aim of follow-up is to determine whether participants developed the outcome under study, although most cohort studies also collect additional information on exposure and confounders to determine changes in exposure status (e.g., a smoker who quits smoking) and other relevant outcomes (e.g., development of other diseases or death). As with exposures, the method of collecting information on outcomes depends on the type of outcome and the degree of desired diagnostic accuracy. Often, mailed questionnaires and phone interviews are used to track participants and determine whether they developed the disease under study. For certain types of outcome (e.g., death or development of cancer), existing vital records (e.g., the National Death Index in the United States) or cancer registration databases could be used to identify study participants who died or developed cancer. Sometimes, in-person interviews and clinic visits are required to accurately determine whether a participant has developed the outcome, as in the case of studies of the incidence of often asymptomatic diseases such as hypertension or HIV infection.

In certain designs, called *repeated measurements designs,* the above measurements are performed more than once for each participant. Examples include *pre-post exposure studies,* in which an assessment such as blood pressure measurement is made before and after an intervention such as the administration of an antihypertensive

medication. Pre-post designs are more commonly used in experimental studies or clinical trials, but there are occasions where this design can be used in observational cohort studies. For instance, results of hearing tests performed during routine preemployment medical examinations can be compared with results from hearing tests performed after a certain period of employment to assess the effect of working in a noisy workplace on hearing acuity.

In *longitudinal repeated measurements designs,* typically two or more exposure (and outcome) measurements are performed over time. These studies tend to be observational and could therefore be carried out prospectively or less commonly retrospectively using precollected data. These studies are ideal for the study of complex phenomena such as the natural history of chronic diseases, including cancer. The repeated measurements allow the investigator to relate changes in time-dependent exposures to the dynamic status of the disease or condition under study. This is especially valuable if exposures are transient and may not be measurable by the time the disease is detected. For instance, repeated measurements are often used in longitudinal studies to examine the natural history of cervical cancer as it relates to infections with the human papilloma virus. In such studies, participants are typically followed up for years with prescheduled clinic visits at certain intervals (e.g., every 6 months). At each visit, participants are examined for evidence of infection with human papilloma virus or development of cervical cancer. Because of the frequent testing for these conditions, it is possible to acquire a deeper understanding of the complex sequence of events that terminates with the development of cancer.

One important goal in all cohort studies is to minimize voluntary loss to follow-up due to participants' dropping out of the study or due to researchers' failure to locate and contact all participants. The longer the study takes to complete, the more likely that a significant proportion of the study participants will be lost to follow-up because of voluntary or involuntary reasons (e.g., death, migration, or development of other diseases). Regardless of the reasons, loss to follow-up is costly because it reduces the study's sample size and therefore its statistical power. More important, loss to follow-up could bias the study results if the remaining cohort members differ from those who were lost to follow-up with respect to the exposure under study. For instance, in a study of the effects of smoking on dementia, smoking may misleadingly appear to reduce the risk of dementia because smokers are more likely than nonsmokers to die at younger age, before dementia could be diagnosed.

## Analysis of Data

Compared with other observational epidemiologic designs, cohort studies provide data permitting the calculations of several types of disease occurrence measures, including disease prevalence, incidence, and cumulative incidence. Typically, disease incidence rates are calculated separately for each of the exposed and the unexposed study groups. The ratio between these rates, the *rate ratio,* is then used to estimate the degree of increased risk of the disease due to the exposure.

In practice, more sophisticated statistical methods are needed to account for the lack of randomization. These methods include direct or indirect standardization, commonly used to account for differences in the age or gender composition of the exposed and unexposed groups. *Poisson regression* could be used to account for differences in one or more confounders. Alternatively, life table and other survival analysis methods, including *Cox proportional hazard models,* could be used to analyze data from cohort studies.

## Special Types

### *Nested Case–Control Studies*

In a nested case–control study, cases and controls are sampled from a preexisting and usually well-defined cohort. Typically, all subjects who develop the outcome under study during follow-up are included as cases. The investigator then randomly samples a subset of noncases (subjects who did not develop the outcome at the time of diagnosis of cases) as controls. Nested case–control studies are more efficient than cohort studies because exposures are measured only for cases and a subset of noncases rather than for all members of the cohort.

For example, to determine whether a certain hormone plays a role in the development of prostate cancer, one could assemble a cohort of men, measure the level of that hormone in the blood of all members of the cohort, and then follow up with them until development of prostate cancer. Alternatively, a nested case–control design could be employed if one has access to a suitable cohort. In this scenario, the investigator would identify all men in the original cohort who developed prostate cancer during follow-up (the cases). For each case, one or more controls would then be selected from among the men who were still alive and free of prostate cancer at the time of the case's diagnosis. Hormone levels would be then measured for only the identified cases and controls rather than for all members of the cohort.

Nested case–control designs have most of the strengths of prospective cohort studies, including the ability to demonstrate that the exposure occurred before the disease. However, the main advantage of nested case–control designs is economic efficiency. Once the original cohort is established, nested case–control designs require far less financial resources and are quicker in producing results than establishing a new cohort. This is especially the case when outcomes are rare or when the measurement of the exposure is costly in time and resources.

### *Case–Cohort Studies*

Similar to the nested case–control study, in a case–cohort study, cases and controls are sampled from a preexisting and usually well-defined cohort. However, the selection of controls is not dependent on the time spent by the noncases in the cohort. Rather the adjustment for confounding effects of time is performed during analysis. Case–cohort studies are more efficient and economical than full cohort studies, for the same reasons that nested case–control studies are economical. In addition, the same set of controls can be used to study multiple outcomes or diseases.

*Salaheddin M. Mahmud*

***See also*** Case-Only Design; Ethics in the Research Process

**Further Readings**

Donaldson, L., & Donaldson, R. (2003). *Essential public health* (2nd ed.). Newbury, UK: LibraPharm.

Rothman, K. J., & Greenland, S. (1998). *Modern epidemiology* (2nd ed.). Philapelphia: Lippincott-Raven.

Singh, H., & Mahmud, S. M. (2009). Different study designs in the epidemiology of cancer: Case-control vs. cohort studies. *Methods in Molecular Biology, 471,* 217–225.

## COLLINEARITY

Collinearity is a situation in which the predictor, or exogenous, variables in a linear regression model are linearly related among themselves or with the intercept term, and this relation may lead to adverse effects on the estimated model parameters, particularly the regression coefficients and their associated standard errors. In practice, researchers often treat correlation between predictor variables as collinearity, but strictly speaking they are not the same; strong correlation implies collinearity, but the opposite is not necessarily true. When there is strong collinearity in a linear regression model, the model estimation procedure is not able to uniquely identify the regression coefficients for highly correlated variables or terms and therefore cannot separate the covariate effects. This lack of identifiability affects the interpretability of the regression model coefficients, which can cause misleading conclusions about the relationships between variables under study in the model.

### Consequences of Collinearity

There are several negative effects of strong collinearity on estimated regression model parameters that can interfere with inference on the relationships of predictor variables with the response variable in a linear regression model. First, the interpretation of regression coefficients as marginal effects is invalid with strong collinearity, as it is not possible to hold highly correlated variables constant while increasing another correlated variable one unit. In addition, collinearity can make regression coefficients unstable and very sensitive to change. This is typically manifested in a large

change in magnitude or even a reversal in sign in one regression coefficient after another predictor variable is added to the model or specific observations are excluded from the model. It is especially important for inference that a possible consequence of collinearity is a sign for a regression coefficient that is counterintuitive or counter to previous research. The instability of estimates is also realized in very large or inflated standard errors of the regression coefficients. The fact that these inflated standard errors are used in significance tests of the regression coefficients leads to conclusions of insignificance of regression coefficients, even, at times, in the case of important predictor variables. In contrast to inference on the regression coefficients, collinearity does not impact the overall fit of the model to the observed response variable data.

## Diagnosing Collinearity

There are several commonly used exploratory tools to diagnose potential collinearity in a regression model. The numerical instabilities in analysis caused by collinearity among regression model variables lead to correlation between the estimated regression coefficients, so some techniques assess the level of correlation in both the predictor variables and the coefficients. Coefficients of correlation between pairs of predictor variables are statistical measures of the strength of association between variables. *Scatterplots* of the values of pairs of predictor variables provide a visual description of the correlation among variables, and these tools are used frequently. There are, however, more direct ways to assess collinearity in a regression model by inspecting the model output itself. One way to do so is through coefficients of correlation of pairs of estimated regression coefficients. These statistical summary measures allow one to assess the level of correlation among different pairs of covariate effects as well as the correlation between covariate effects and the intercept. Another way to diagnose collinearity is through *variance inflation factors,* which measure the amount of increase in the estimated variances of regression coefficients compared with when predictor variables are uncorrelated. Drawbacks of the variance inflation factor as a collinearity diagnostic tool are that it does not illuminate the nature of the collinearity, which is problematic if the collinearity is between more than two variables, and it does not consider collinearity with the intercept. A diagnostic tool that accounts for these issues consists of variance-decomposition proportions of the regression coefficient variance–covariance matrix and the condition index of the matrix of the predictor variables and constant term. Some less formal diagnostics of collinearity that are commonly used are a counterintuitive sign in a regression coefficient, a relatively large change in value for a regression coefficient after another predictor variable is added to the model, and a relatively large standard error for a regression coefficient. Given that statistical inference on regression coefficients is typically a primary concern in regression analysis, it is important for one to apply diagnostic tools in a regression analysis before interpreting the regression coefficients, as the effects of collinearity could go unnoticed without a proper diagnostic analysis.

## Remedial Methods for Collinearity

There are several methods in statistics that attempt to overcome collinearity in standard linear regression models. These methods include *principal components regression, ridge regression,* and a technique called the *lasso*. Principal components regression is a variable subset selection method that uses combinations of the exogenous variables in the model, and ridge regression and the lasso are penalization methods that add a constraint on the magnitude of the regression coefficients. Ridge regression was designed precisely to reduce collinearity effects by penalizing the size of regression coefficients. The lasso also shrinks regression coefficients, but it shrinks the least significant variable coefficients toward zero to remove some terms from the model. Ridge regression and the lasso are considered superior to principal components regression to deal with collinearity in regression models because they more purposely reduce inflated variance in regression coefficients due to collinearity while retaining interpretability of individual covariate effects.

### Future Research

While linear regression models offer the potential for explaining the nature of relationships between variables in a study, collinearity in the exogenous variables and intercept can generate curious results in the estimated regression coefficients that may lead to incorrect substantive conclusions about the relationships of interest. More research is needed regarding the level of collinearity that is acceptable in a regression model before statistical inference is unduly influenced.

*David C. Wheeler*

***See also*** Coefficients of Correlation, Alienation, and Determination; Correlation; Covariate; Exogenous Variable; Inference: Deductive and Inductive; Parameters; Predictor Variable; Regression Coefficient; Significance, Statistical; Standard Error of Estimate; Variance

#### Further Readings

Belsley, D. A. (1991). *Conditioning diagnostics: Collinearity and weak data in regression*. New York: Wiley.

Fox, J. (1997). *Applied regression analysis, linear models, and related methods*. Thousand Oaks, CA: Sage.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer-Verlag.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear regression models*. Chicago: Irwin.

## Column Graph

A column graph summarizes categorical data by presenting parallel vertical bars with a height (and hence area) proportionate to specific quantities of data for each category. This type of graph can be useful in comparing two or more distributions of nominal- or ordinal-level data.

### Developing Column Graphs

A column graph can and should provide an easy-to-interpret visual representation of a frequency or percentage distribution of a single (or multiple) variable(s). Column graphs present a series of vertical equal-width rectangles, each with a height proportional to the frequency (or percentage) of a specific category of observations. Categories are labeled on the x-axis (the horizontal axis), and frequencies are labeled on the y-axis (the vertical axis). For example, a column graph that displays the partisan distribution of a single session of a state legislature would consist of at least two rectangles, one representing the number of Democratic seats, one representing the number of Republican seats, and perhaps one representing the number of independent seats. The x-axis would consist of the "Democratic" and "Republican" labels while the y-axis would consist of labels representing intervals for the number of seats in the state legislature.

When developing a column graph, it is vital that the researcher present a set of categories that is both exhaustive and mutually exclusive. In other words, each potential value must belong to one and only one category. Developing such a category schema is relatively easy when the researcher is faced with discrete data, in which the number observations for a particular value can be counted (i.e. nominal and ordinal level data). For nominal data, the categories are unordered, with interchangeable values (e.g., gender, ethnicity, religion). For ordinal data, the categories exhibit some type of relation to each other, although the relation does not exhibit specificity beyond ranking the values (e.g., greater than vs. less than, *agree strongly* vs. *agree* vs. *disagree* vs. *disagree strongly*). Because nominal and ordinal data are readily countable, once the data are obtained, the counts can then be readily transformed into a column graph.

With continuous data (i.e., interval- and ratio-level data), an additional step is required. For interval data, the distance between observations is fixed (e.g., Carolyn makes $5,000 more per year than John does). For ratio-level data, the data can be measured as in relation to a fixed point (e.g., a family's income falls a certain percentage above or below the poverty line). For continuous data, the number of potential values can be infinite or, at the very least, extraordinarily high, thus producing a cluttered chart. Therefore, it is best to reduce interval- or ratio-level data to ordinal-level data by collapsing the range of potential values into a few select categories. For example, a survey that asks

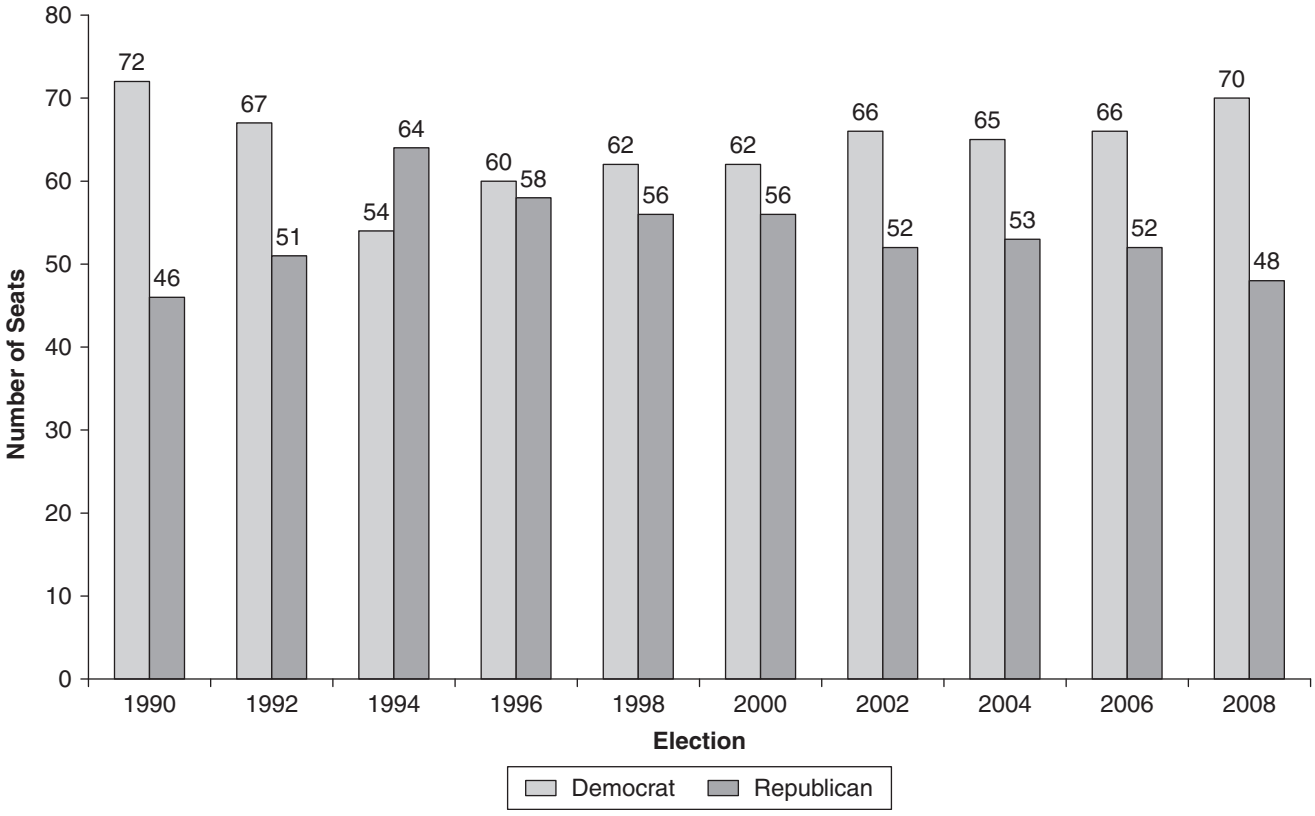



**Figure 1** 1990–2009 Illinois House Partisan Distribution (Column Graph)

*Source: Almanac of Illinois Politics*. (2009). Springfield: Illinois Issues.

respondents for their age could produce dozens of potential answers, and therefore it is best to condense the variable into a select few categories (e.g., 18–25, 26–35, 36–45, 46–64, 65 and older) before making a column graph that summarizes the distribution.

## Multiple Distribution Column Graphs

Column graphs can also be used to compare multiple distributions of data. Rather than presenting a single set of vertical rectangles that represents a single distribution of data, column graphs present multiple sets of rectangles, one for each distribution. For ease of interpretation, each set of rectangles should be grouped together and separate from the other distributions. This type of graph can be particularly useful in comparing counts of observations across different categories of interest. For example, a researcher conducting a survey might present the aggregate distribution of self-reported partisanship, but the researcher can also demonstrate the gender gap by displaying separate partisan distributions for male and female respondents. By putting each distribution into a single graph, the researcher can visually present the gender gap in a readily understandable format.

Column graphs might also be used to explore chronological trends in distributions. One such example is in Figure 1, which displays the partisan makeup of the Illinois state House from the 1990 through the 2008 elections. The black bars represent the number of Republican seats won in the previous election (presented on the x-axis) and the gray bars represent the number of Democratic seats. By showing groupings of the two partisan bars side by side, the graph provides for an easy interpretation of which party had control of the legislative chamber after a given election. Furthermore, readers can look across each set of bars to

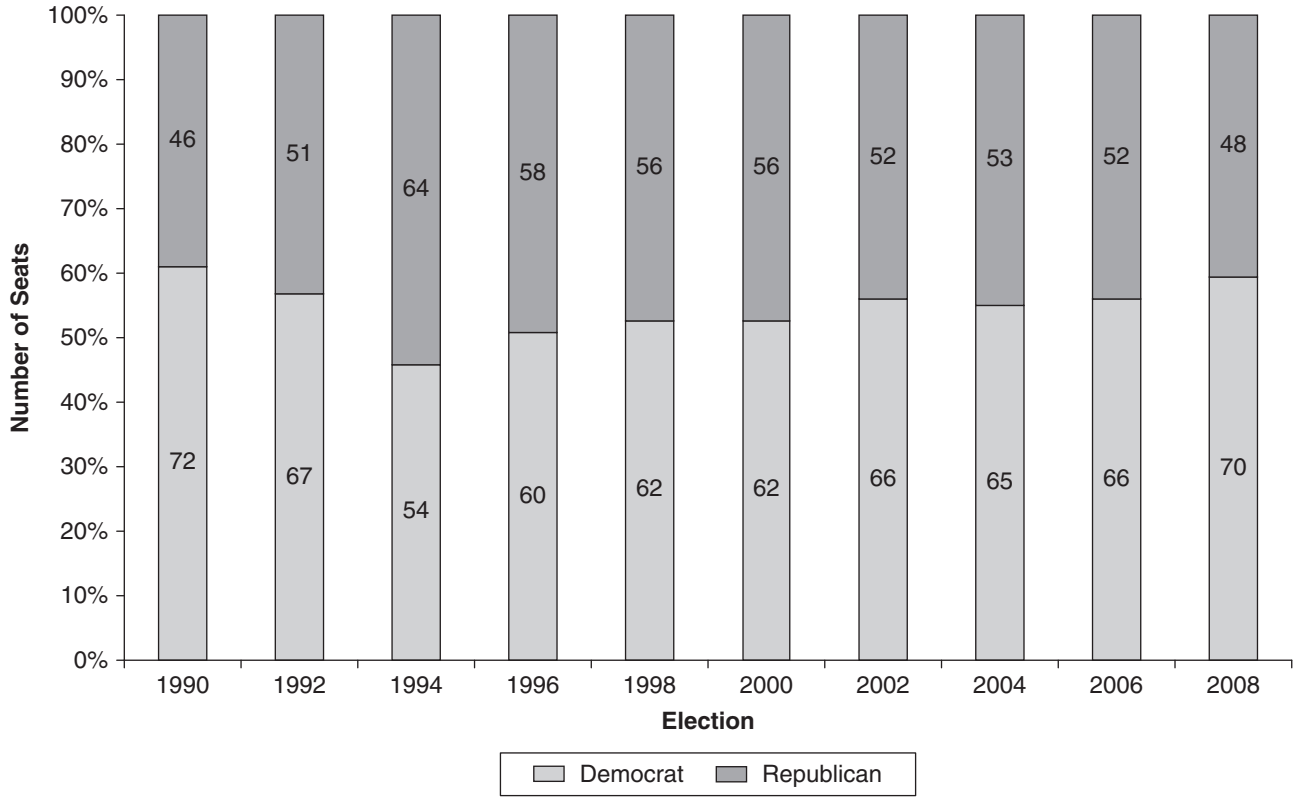



**Figure 2** 1990–2009 Illinois House Partisan Distribution (100% Stacked Column Graph)

*Source: Almanac of Illinois Politics.* (2009). Springfield: Illinois Issues.

determine the extent to which the partisan makeup of the Illinois House has changed over time.

### *Stacked Column*

One special form of the column graph is the stacked column, which presents data from a particular distribution in a single column. A regular stacked column places the observation counts directly on top of one another, while a 100% stacked column does the same while modifying each observation count into a percentage of the overall distribution of observations. The former type of graph might be useful when the researcher wishes to present a specific category of interest (which could be at the bottom of the stack) and the total number of observations. The latter might be of interest when the researcher wants a visual representation of how much of the total distribution is represented by each value and the extent to which that distribution is affected by other variables of interest (such as time). Either version of the stacked column approach can be useful in comparing multiple distributions, such as with Figure 2, which presents a 100% stacked column representation of chronological trends in Illinois House partisan makeup.

## Column Graphs and Other Figures

When one is determining whether to use a column graph or some other type of visual representation of relevant data, it is important to consider the main features of other types of figures, particularly the level of measurement and the number of distributions of interest.

### *Bar Graph*

A column graph is a specific type of bar graph (or chart). Whereas the bar graph can present summaries of categorical data in either vertical or

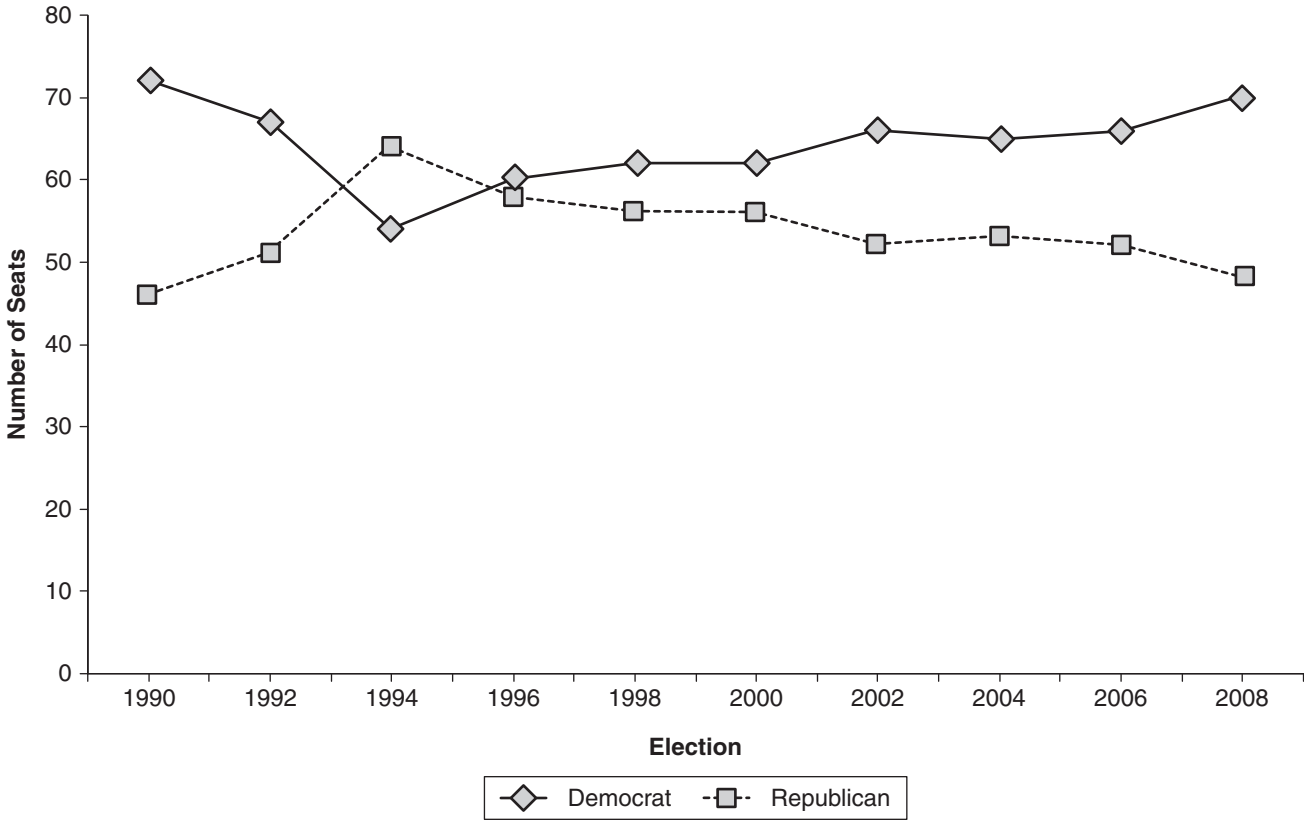



**Figure 3** 1990–2009 Illinois House Partisan Distribution (Line Graph)

*Source: Almanac of Illinois Politics*. (2009). Springfield: Illinois Issues.

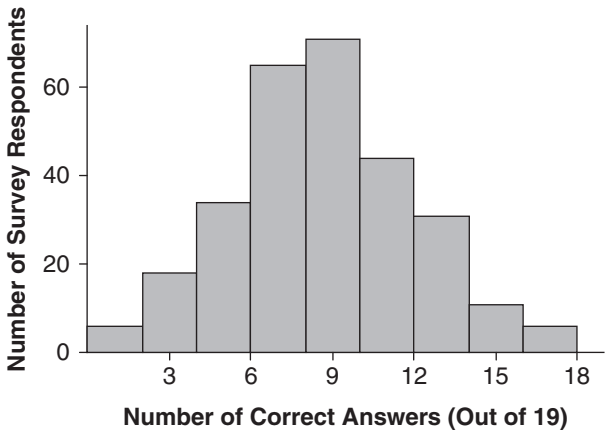



**Figure 4** Political Knowledge Variable Distribution (Histogram)

*Source:* Adapted from the University of Illinois Subject Pool, Bureau of Educational Research.

horizontal rectangles, a column graph presents only vertical rectangles. The decision to present horizontal or vertical representations of the data is purely aesthetic, and so the researcher might want to consider which approach would be the most visually appealing, given the type of analysis presented by the data.

### *Pie Chart*

A pie chart presents the percentage of each category of a distribution as a segment of a circle. This type of graph allows for only a single distribution at a time; multiple distributions require multiple pie charts. As with column and bar graphs, pie charts represent observation counts (or percentages) and thus are used for discrete data, or at the

very least continuous data collapsed into a select few discrete categories.

### *Line Graph*

A line graph displays relationships between two changing variables by drawing a line that connects actual or projected values of a dependent variable (y-axis) based on the value of an independent variable (x-axis). Because line graphs display trends between plots of observations across an independent variable, neither the dependent nor the independent variable can contain nominal data. Furthermore, ordinal data do not lend themselves to line graphs either, because the data are ordered only within the framework of the variables. As with column graphs and bar graphs, line graphs can track multiple distributions of data based on categories of a nominal or ordinal variable. Figure 3 provides such an example, with yet another method of graphically displaying the Illinois House partisan makeup. The election year serves as the independent variable, and the number of Illinois House seats for a particular party as the dependent variable. The solid gray and dotted black lines present the respective Democratic and Republican legislative seat counts, allowing for easy interpretation of the partisan distribution trends.

### *Histogram*

A histogram is a special type of column graph that allows for a visual representation of a single frequency distribution of interval or ratio data without collapsing the data to a few select categories. Visually, a histogram looks similar to a column graph, but without any spaces between the rectangles. On the x-axis, a histogram displays intervals rather than discrete categories. Unlike a bar graph or a column graph, a histogram only displays distributions and cannot be used to compare multiple distributions. The histogram in Figure 4 displays the distribution of a political knowledge variable obtained via a 2004 survey of University of Illinois undergraduate students. The survey included a series of 19 questions about U.S. government and politics.

*Michael A. Lewkowicz*

***See also*** Bar Chart; Histogram; Interval Scale; Line Graph; Nominal Scale; Ordinal Scale; Pie Chart; Ratio Scale

#### Further Readings

Frankfort-Nachmias, C., & Nachmias, D. (2007). *Research methods in the social sciences* (7th ed.). New York: Worth.

Harris, R. L. (2000). *Information graphics: A comprehensive illustrated reference*. New York: Oxford University Press.

Stoval, J. G. (1997). *Infographics: A journalist's guide*. Boston: Allyn and Bacon.

## Completely Randomized Design

A completely randomized design (CRD) is the simplest design for comparative experiments, as it uses only two basic principles of experimental designs: *randomization* and *replication*. Its power is best understood in the context of agricultural experiments (for which it was initially developed), and it will be discussed from that perspective, but true experimental designs, where feasible, are useful in the social sciences and in medical experiments.

In CRDs, the treatments are allocated to the experimental units or plots in a completely random manner. CRD may be used for single- or multifactor experiments. This entry discusses the application, advantages, and disadvantages of CRD studies and the processes of conducting and analyzing them.

### Application

CRD is mostly useful in laboratory and green house experiments in agricultural, biological, animal, environmental, and food sciences, where experimental material is reasonably homogeneous. It is more difficult when the experimental units are people.

### Advantages and Disadvantages

This design has several advantages. It is very flexible as any number of treatments may be used, with equal or unequal replications. The design has a comparatively simple statistical analysis and

retains this simplicity even if some observations are missing or lost accidentally. The design provides maximum degrees of freedom for the estimation of error variance, which increases the precision of an experiment.

However, the design is not suitable if a large number of treatments are used and the experimental material is not reasonably homogeneous. Therefore, it is seldom used in agricultural field experiments in which soil heterogeneity may be present because of soil fertility gradient or in animal sciences when the animals (experimental units) vary in such things as age, breed, or initial body weight, or with people.

## Layout of the Design

The plan of allocation of the treatments to the experimental material is called the layout of the design.

Let the $i$th ($i = 1, 2, \ldots, v$) treatments be replicated $r_i$ times. Therefore, $N = \sum r_i$ is the total number of required experimental units.

The treatments are allocated to the experimental units or plots in a completely random manner. Each treatment has equal probability of allocation to an experimental unit.

Given below is layout plan of CRD with four treatments, denoted by integers, each replicated 3 times and allocated to 12 experimental units.

| | | | |
|---|---|---|---|
| 3 | 2 | 4 | 4 |
| 3 | 1 | 4 | 1 |
| 3 | 1 | 2 | 2 |

## Randomization

Some common methods of random allocation of treatments to the experimental units are illustrated in the following:

Consider an experiment with less than or up to 10 treatments. In this case, a 1-digit random number table can be used. The treatments are allotted a number each. The researcher picks up random numbers with replacement (i.e., a random number may get repeated) from the random number table until the number of replications of that treatment is exhausted.

For experiments with more than 10 treatments, a 2-digit random number table or a combination of two rows or columns of 1-digit random numbers can be used. Here each 2-digit random number is divided by the number of treatments, and the residual number is selected. When the residual is 00, the divisor number is selected.The digit 00 already occurring in the table is discarded. The digit 00 is discarded.

On small pieces of paper identical in shape and size, the numbers $1, 2, \ldots, N$ are written. These are thoroughly mixed in a box, the papers are drawn one by one, and the numbers on the selected papers are random numbers. After each draw, the piece of paper is put back, and thorough mixing is performed again.

The random numbers may also be generated by computers.

## Statistical Analysis

The data from CRD are examined by the method of one-way analysis of variance (ANOVA) for classification.

The analysis of data from CRD is performed using the linear fixed-effects model given as

$$y_{ij} = \mu + t_i + \varepsilon_{ij}, i = 1, 2, \ldots, v; j = 1, 2, \ldots, r_i;$$

where $y_{ij}$ is the yield or response from the $j$th unit receiving the $i$th treatment, $t_i$ is the fixed effect of the $i$th treatment, and $\varepsilon_{ij}$ is random error effect. Suppose we are interested in knowing whether wheat can be grown in a particular agroclimatic situation. Then we will randomly select some varieties of wheat, from all possible varieties, for testing in an experiment. Here the effect of treatment is random. When the varieties are the only ones to be tested in an experiment, the effect of varieties (treatments) is fixed.

The experimental error is a random variable with mean zero and constant variance($\sigma^2$). The experimental error is normally and independently distributed with mean zero and constant variance for every treatment.

Let $N = \sum r_i$ be the total number of experimental units, $\sum\sum y_{ij} = y\ldots = G =$ grand total of all the observations, and $\sum y_{ij} = y = T_i =$ the total response from the experimental units receiving the ith treatment. Then,

**Table 1** Analysis of Variance for Completely Randomized Design

| *Sv* | *df* | *SS* | *MSS* | Variance Ratio |
|---|---|---|---|---|
| Between treatments | $v-1$ | *TrSS* | $s_t^2 = TrSS/v-1$ | $F_t = s_t^2/s_e^2$ |
| Within treatments | $N-v$ | *ErSS* | $s_e^2 = ErSS/N-v$ | |

*Note:* *Sv* sources of variation; *df* = degrees of freedom; *SS* = sum of squares; *MSS* = mean sum of squares (variance); *TrSS* = treatments sum of squares; *ErSS* = error sum of squares

$$\sum\sum(y_{ij}-\mu)^2 = \sum r_i(y_i\ldots-\overline{y}\ldots)^2 + \sum\sum(y_{ij}-\overline{y}_i.)^2, i = 1,\ldots,v; j = 1,2,\ldots,r_i,$$

that is, the total sum of squares is equal to the sum of the treatments sum of squares and the error sum of squares. This implies that the total variation can be partitioned into two components: between treatments and within treatments (experimental error).

Table 1 is an ANOVA table for CRD.

Under the null hypothesis, $H_0 = t_1 = t_2 = \cdots = t_v$; that is, all treatment means are equal, the statistic $F_t = s_t^2/s_e^2$ follows the $F$ distribution with $(v-1)$ and $(N-v)$ degrees of freedom.

If $F$ (observed) $\geq F$ (expected) for the same *df* and at a specified level of significance, say $\alpha$%, then the null hypothesis of equality of means is rejected at $\alpha$% level of significance.

If $H_0$ is rejected, then the pairwise comparison between the treatment means is made by using the critical difference (CD) test,

$$CD\alpha\% = SE_d \cdot t \text{ for error } df \text{ and } \alpha\% \text{ level of significance,}$$

where $SE_d = \sqrt{ErSS(1/r_i + 1/r_j)/(N-v)}$, $SE_d$ stands for standard error of difference between means of two treatments and when $r_i = r_j, SE_d = \sqrt{2ErSS/\{r(N-v)\}}$.

If the difference between any two treatment means is greater than or equal to the critical difference, the treatment means are said to differ significantly.

The critical difference is also called the *least significant difference.*

## Number of Replications

The number of replications required for an experiment is affected by the inherent variability and size of the experimental material, the number of treatments, and the degree of precision required. The minimum number of replications required to detect the specified differences between two treatment means at a specified level of significance is given by the $t$ statistic at error *df* and $\alpha$% level of significance, $t = d/\{s_e\sqrt{2/r}\}$, where $d = \overline{X}_1 - \overline{X}_2$, $\overline{X}_1$ and $\overline{X}_2$ are treatment means with each treatment replicated $r$ times, and $s_e^2$ is error variance.

Therefore,

$$r = 2t^2s_e^2/d^2.$$

It is observed that from 12 *df* onwards, the values of $t$ and $F$ (for smaller error variance) decrease considerably slowly, and so from empirical considerations the number of replications are so chosen as to provide about 12 *df* for error variance for the experiment.

## Nonconformity to the Assumptions

When the assumptions are not realized, the researcher may apply one of the various transformations in order to bring improvements in assumptions. The researcher then may proceed with the usual ANOVA after the transformation. Some common types of transformations are described below.

### *Arc Sine Transformation*

This transformation, also called *angular transformation,* is used for count data obtained from binomial distribution, such as success–failure, diseased–nondiseased, infested–noninfested, barren–nonbarren tillers, inoculated–noninoculated, male–female, dead–alive, and so forth.

The transformation is not applicable to percentages of carbohydrates, protein, profit, disease index, and so forth, which are not derived from count data. The transformation is not needed if nearly all the data lie between 30%

and 70%. The transformation may be used when data range from 0% to 30% and beyond or from 100% to 70% and below. For transformation, $0/n$ and $n/n$ should be taken as $1/4n$ and $(1 - 1/4n)$, respectively.

### *Square Root Transformation*

The square root transformation is used for data from a Poisson distribution, that is, when data are counts of rare events such as number of defects or accidents, number of infested plants in a plot, insects caught on traps, or weeds per plot. The transformation consists of taking the square root of each observation before proceeding with the ANOVA in the usual manner.

### *Logarithmic Transformation*

The logarithmic transformation is used when standard deviation is proportional to treatment means, that is, if the coefficient of variation is constant. The transformation achieves additivity. It is used for count data (large whole numbers) covering a wide range, such as number of insects per plant or number of egg masses. When zero occurs, 1 is added to it before transformation.

*Kishore Sinha*

*See also* Analysis of Variance (ANOVA); Experimental Design; Fixed-Effects Models; Randomized Block Design; Research Design Principles

#### Further Readings

Bailey, R. A. (2008). *Design of comparative experiments*. Cambridge, UK: Cambridge University Press.
Box, G. E. P., Hunter, W. G., & Hunter, J. S. (1978). *Statistics for experiments*. New York: Wiley.
Dean, A., & Voss, D. (1999). *Design and analysis of experiments*. New York: Springer.
Montgomery, D. C. (2001). *Design and analysis of experiments*. New York: Wiley.

## Computerized Adaptive Testing

Adaptive testing, in general terms, is an assessment process in which the test items administered to examinees differ based on the examinees' responses to previous questions. Computerized adaptive testing uses computers to facilitate the "adaptive" aspects of the process and to automate scoring. This entry discusses historical perspectives, goals, psychometric and item selection approaches, and issues associated with adaptive testing in general and computerized adaptive testing in particular.

### Historical Perspectives

Adaptive testing is not new. Through the ages examiners have asked questions and, depending on the response given, have chosen different directions for further questioning for different examinees. Clinicians have long taken adaptive approaches, and so since the advent of standardized intelligence testing, many such tests have used adaptive techniques. For both the 1916 edition of the Stanford-Binet and the 1939 edition of the Wechsler-Bellevue intelligence tests, the examiner chose a starting point and, if the examinee answered correctly, asked harder questions until a string of incorrect answers was provided. If the first answer was incorrect, an easier starting point was chosen.

Early group-administered adaptive tests faced administration problems in the precomputerized administration era. Scoring each item and making continual routing decisions was too slow a process. One alternative explored by William Angoff and Edith Huddleston in the 1950s was two-stage testing. An examinee would take a half-length test of medium difficulty, that test was scored, and then the student would be routed to either an easier or a harder last half of the test. Another alternative involved use of a special marker that revealed invisible ink when used. Examinees would use their markers to indicate their responses. The marker would reveal the number of the next item they should answer. These and other approaches were too complex logistically and never became popular.

Group-administered tests could not be administered adaptively with the necessary efficiency until the increasing availability of computer-based testing systems, circa 1970. At that time research on computer-based testing proliferated. In 1974, Frederic Lord (inventor of the theoretical underpinnings of

much of modern psychometric theory) suggested the field would benefit from researchers' getting together to share their ideas. David Weiss, a professor at the University of Minnesota, thought this a good idea and coordinated a series of three conferences on computerized adaptive testing in 1975, 1977, and 1979, bringing together the greatest thinkers in the field. These conferences energized the research community, which focused on the theoretical underpinnings and research necessary to develop and establish the psychometric quality of computerized adaptive tests.

The first large-scale computerized adaptive tests appeared circa 1985, including the U.S. Army's Computerized Adaptive Screening Test, the College Board's Computerized Placement Tests (the forerunner of today's Accuplacer), and the Computerized Adaptive Differential Ability Tests of the Psychological Corporation (now part of Pearson Education). Since that time computerized adaptive tests have proliferated in all spheres of assessment.

## Goals of Adaptive Testing

The choice of test content and administration mode should be based on the needs of a testing program. What is best for one program is not necessarily of importance for other programs. There are primarily three different needs that can be addressed well by adaptive testing: maximization of test reliability for a given testing time, minimization of individual testing time to achieve a particular reliability or decision accuracy, and the improvement of diagnostic information.

### *Maximization of Test Reliability*

Maximizing reliability for a given testing time is possible because in conventional testing, many examinees spend time responding to items that are either trivially easy or extremely hard, and thus many items do not contribute to our understanding of what the examinee knows and can do. If a student can answer complex multiplication and division questions, there is little to be gained asking questions about single-digit addition. When multiple-choice items are used, students are sometimes correct when they randomly guess the answers to very difficult items. Thus the answers to items that are very difficult for a student are worse than useless—they can be misleading.

An adaptive test maximizes reliability by replacing items that are too difficult or too easy with items of an appropriate difficulty. Typically this is done by ordering items by difficulty (usually using sophisticated statistical models such as *Rasch scaling* or the more general *item response theory*) and administering more difficult items subsequent to correct responses and easier items after incorrect responses.

By tailoring the difficulty of the items administered dependent on prior examinee responses, a well-developed adaptive test can achieve the reliability of a conventional test with approximately one half to one third the items or, alternatively, can achieve a significantly higher reliability in a given amount of time. (However, testing time is unlikely to be reduced in proportion to the item reduction because very easy and very difficult items often require little time to answer correctly, or not try, respectively.)

### *Minimization of Testing Time*

An alternative goal that can be addressed with adaptive testing is minimizing the testing time needed to achieve a fixed level of reliability. Items can be administered until the reliability of the test for a particular student, or the decision accuracy for a test that classifies examinees into groups (such as pass–fail), reaches an acceptable level. If the match between the proficiency of an examinee and the quality of the item pool is good (e.g., if there are many items that measure proficiency particularly well within a certain range of scores), few items will be required to determine an examinee's proficiency level with acceptable precision. Also, some examinees may be very consistent in their responses to items above or below their proficiency level—it might require fewer items to pinpoint their proficiency compared with people who are inconsistent.

Some approaches to adaptive testing depend on the assumption that items can be ordered along a single continuum of difficulty (also referred to as *unidimensionality*). One way to look at this is that the ordering of item difficulties is the same for all

identifiable groups of examinees. In some areas of testing, this is not a reasonable assumption. For example, in a national test of science for eighth-grade students, some students may have just studied life sciences and others may have completed a course in physical sciences. On average, life science questions will be easier for the first group than for the second, while physical science questions will be relatively easier for the second group. Sometimes the causes of differences in dimensionality may be more subtle, and so in many branches of testing, it is desirable not only to generate an overall total score but to produce a diagnostic profile of areas of strength and weakness, perhaps with the ultimate goal of pinpointing remediation efforts.

An examinee who had recently studied physical science might get wrong a life science item that students in general found easy. Subsequent items would be easier, and that student might never receive the relatively difficult items that she or he could have answered correctly. That examinee might receive an underestimate of her or his true level of science proficiency.

One way to address this issue for multidimensional constructs would be to test each such construct separately. Another alternative would be to create groups of items balanced on content—*testlets*—and give examinees easier or harder testlets as the examinees progress through the test.

### *Improvement of Diagnostic Information*

Diagnostic approaches to computerized adaptive testing are primarily in the talking stage. Whereas one approach to computerized adaptive diagnostic testing would be to use multidimensional item response models and choose items based on the correctness of prior item responses connected to the difficulty of the items on the multiple dimensions, an alternative approach is to lay out a tested developmental progression of skills and select items based on precursor and "postcursor" skills rather than on simple item difficulty. For example, if one posits that successful addition of two-digit numbers with carrying requires the ability to add one-digit numbers with carrying, and an examinee answers a two-digit addition with carrying problem incorrectly, then the next problem asked should be a one-digit addition with carrying problem. Thus, in such an approach, expert knowledge of relationships within the content domain drives the selection of the next item to be administered. Although this is an area of great research interest, at this time there are no large-scale adaptive testing systems that use this approach.

## Psychometric Approaches to Scoring

There are four primary psychometric approaches that have been used to support item selection and scoring in adaptive testing: maximum likelihood item response theory, Bayesian item response theory, classical test theory, and decision theory.

*Maximum likelihood item response theory* methods have long been used to estimate examinee proficiency. Whether they use the Rasch model, three-parameter logistic model, partial credit, or other item response theory models, fixed (previously estimated) item parameters make it fairly easy to estimate the proficiency level most likely to have led to the observed pattern of item scores.

*Bayesian methods* use information in addition to the examinee's pattern of responses and previously estimated item parameters to estimate examinee proficiency. Bayesian methods assume a population distribution of proficiency scores, often referred to as a *prior,* and use the unlikeness of an extreme score to moderate the estimate of the proficiency of examinees who achieved such extreme scores. Since examinees who have large positive or negative errors of measurement tend to get more extreme scores, Bayesian approaches tend to be more accurate (have smaller errors of measurement) than maximum likelihood approaches.

The essence of Bayesian methods for scoring is that the probabilities associated with the proficiency likelihood function are multiplied by the probabilities in the prior distribution, leading to a new, posterior, distribution. Once that posterior distribution is produced, there are two primary methods for choosing an examinee score. The first is called *modal estimation* or *maximum a posteriori*—pick the proficiency level that has the highest probability (probability density). The second method is called *expected a poste-riori* and is essentially the mean of the posterior distribution.

**Table 1** Record for Two Hypothetical Examinees Taking a Mastery Test Requiring 60% Correct on the Full-Length (100-item) Test

| | | *Examinee 1* | | *Examinee 2* | |
|---|---|---|---|---|---|
| *After Testlet Number* | *Confidence Interval* | *Number Correct* | *Percent Correct* | *Number Correct* | *Percent Correct* |
| 1 | .20–1.00 | 6 | .60 | 9 | .90 |
| 2 | .34–.86 | 11 | .55 | **19** | **.95** |
| 3 | .40–.80 | 18 | .60 | | |
| 4 | .34–.76 | 25 | .63 | | |
| 5 | .37–.73 | 29 | .58 | | |
| 6 | .50–.60 | 34 | .57 | | |
| 7 | .52–.68 | 39 | .56 | | |
| 8 | .54–.66 | 43 | .54 | | |
| 9 | .56–.64 | **48** | **.53** | | |
| 10 | .60–.60 | | | | |

Item response theory proficiency estimates have different psychometric characteristics from *classical test theory* approaches to scoring tests (based on the number of items answered correctly). The specifics of these differences are beyond the scope of this entry, but they have led some testing programs (especially ones in which some examinees take the test traditionally and some take it adaptively) to transform the item response theory proficiency estimate to an estimate of the number right that would have been obtained on a hypothetical base form of the test. From the examinee's item response theory–estimated proficiency and the set of item parameters for the base form items, the probability of a correct response can be estimated for each item, and the sum of those probabilities is used to estimate the number of items the examinee would have answered correctly.

*Decision theory* is a very different approach that can be used for tests that classify or categorize examinees, for example into two groups, passing and failing. One approach that has been implemented for at least one national certification examination was developed by Charles Lewis and Kathleen Sheehan. Parallel testlets—sets of perhaps 10 items that cover the same content and are of about the same difficulty—were developed. After each testlet is administered, a decision is made to either stop testing and make a pass–fail decision or administer another testlet. The decision is based on whether the confidence interval around an examinee's estimated score is entirely above or entirely below the proportion-correct cut score for a pass decision.

For example, consider a situation in which to pass, an examinee must demonstrate mastery of 60% of the items in a 100-item pool. We could ask all 100 items and see what percentage the examinee answers correctly. Alternatively, we could ask questions and stop as soon as the examinee answers 60 items correctly. This could save as much as 40% of computer time at a computerized test administration center. In fact, one could be statistically confident that an examinee will get at least 60% of the items correct administering fewer items.

The Lewis–Sheehan *decision theoretic approach* is based on a branch of statistical theory called *Waldian sequential ratio testing,* in which after each sample of data, the confidence interval is calculated, and a judgment is made, either that enough information is possessed to make a decision or that more data needs to be gathered. Table 1 presents two simplified examples of examinees tested with multiple parallel 10-item testlets. The percent correct is boldfaced when it falls outside of the 99% confidence interval and a pass–fail decision is made.

The confidence interval in this example narrows as the sample of items (number of testlets administered) increases and gets closer to the full 100-item domain. Examinee 1 in Table 1 has a true domain percent correct close to the cut score of .60. Thus

it takes nine testlets to demonstrate the examinee is outside the required confidence interval. Examinee 2 has a true domain percent correct that is much higher than the cut score and thus demonstrates mastery with only two testlets.

## Item Selection Approaches

Item selection can be subdivided into two pieces: selecting the first item and selecting all subsequent items. Selection of the first item is unique because for all other items, some current information about student proficiency (scores from previous items) is available. Several approaches can be used for selecting the first item. Everyone can get the same first item—one that is highly discriminating near the center of the score distribution—but such an approach would quickly make that item nonsecure (i.e., it could be shared easily with future examinees). An alternative would to be to randomly select from a group of items that discriminate well to the center of the score distribution. Yet another approach might be to input some prior information (for example, class grades or previous year's test scores) to help determine a starting point.

Once the first item is administered, several approaches exist for determining which items should be administered next to an examinee. There are two primary approaches to item selection for adaptive tests based on item response theory: maximum information and minimum posterior variance. With *maximum information* approaches, one selects the item that possesses the maximum amount of statistical information at the current estimated proficiency. Maximum information approaches are consistent with maximum likelihood item response theory scoring. *Minimum posterior variance* methods are appropriate for Bayesian scoring approaches. With this approach, the item is selected that, after scoring, will lead to the smallest variance of the posterior distribution, which is used to estimate the examinee's proficiency.

Most adaptive tests do not apply either of these item selection approaches in their purely statistical form. One reason to vary from these approaches is to ensure breadth of content coverage. Consider a physics test covering mechanics, electricity, magnetism, optics, waves, heat, and thermodynamics. If items are chosen solely by difficulty, some examinees might not receive any items about electricity and others might receive no items about optics. Despite items' being scaled on a national sample, any given examinee might have an idiosyncratic pattern of knowledge. Also, the sample of items in the pool might be missing topic coverage in certain difficulty ranges. A decision must be made to focus on item difficulty, content coverage, or some combination of the two.

Another consideration is item exposure. Since most adaptive test item pools are used for extended periods of time (months if not years), it is possible for items to become public knowledge. Sometimes this occurs as a result of concerted cheating efforts, sometimes just because examinees and potential examinees talk to each other.

More complex item selection approaches, such as developed by Wim van der Linden, can be used to construct "on the fly" tests that meet a variety of predefined constraints on item selection.

## Issues

### *Item Pool Requirements*

Item pool requirements for adaptive tests are typically more challenging than for traditional tests. With this proliferation, the psychometric advantages of adaptive testing have often been delivered, but logistical problems have been discovered. When stakes are high (for example, admissions or licensure testing), test security is very important—ease of cheating would invalidate test scores. It is still prohibitively expensive to arrange for a sufficient number of computers in a secure environment to test hundreds of thousands of examinees at one time, as some of the largest testing programs have traditionally done with paper tests. As an alternative, many computerized adaptive testing programs (such as the Graduate Record Examinations) have moved from three to five administrations a year to administrations almost every day. This exposes items to large numbers of examinees who sometimes remember and share the items. Minimizing widespread cheating when exams are offered so frequently requires the creation of multiple item pools, an expensive undertaking, the cost of which is often passed on to examinees.

Another issue relates to the characteristics of the item pool required for adaptive testing. In a traditional norm-referenced test, most items are chosen so that about 60% of the examinees answer the item correctly. This difficulty level maximally differentiates people on a test in which everyone gets the same items. Thus few very difficult items are needed, which is good since it is often difficult to write very difficult items that are neither ambiguous nor trivial. On an adaptive test, the most proficient examinees might all see the same very difficult items, and thus to maintain security, a higher percentage of such items are needed.

### *Comparability*

Comparability of scores on computerized and traditional tests is a very important issue as long as a testing program uses both modes of administration. Many studies have been conducted to determine whether scores on computer-administered tests are comparable to those on traditional paper tests. Results of individual studies have varied, with some implying there is a general advantage to examinees who take tests on computer, others implying the advantage goes to examinees taking tests on paper, and some showing no statistically significant differences. Because different computer-administration systems use different hardware, software, and user interfaces and since these studies have looked at many different content domains, perhaps these different results are real. However, in a 2008 meta-analysis of many such studies of K–12 tests in five subject areas, Neal Kingston showed a mean weighted effect size of $-.01$ standard deviations, which was not statistically significant. However, results differed by data source and thus the particular test administration system used.

### *Differential Access to Computers*

Many people are concerned that differential access to computers might disadvantage some examinees. While the number of studies looking at potential bias related to socioeconomic status, gender, ethnicity, or amount of computer experience is small, most recent studies have not found significant differences in student-age populations.

*Neal Kingston*

*See also* Bayes's Theorem; b Parameter; Decision Rule; Guessing Parameter; Item Response Theory; Reliability; "Sequential Tests of Statistical Hypotheses"

**Further Readings**

Wainer, H. (Ed.). (1990). *Computerized adaptive testing: A primer.* Hillsdale, NJ: Lawrence Erlbaum.
Mills, C., Potenza, M., Fremer, J., & Ward, W. (Ed.) (2002). *Computer-based testing.* Hillsdale, NJ: Lawrence Erlbaum.
Parshall, C., Spray, J., Kalohn, J., & Davey, T. (2002). *Practical considerations in computer-based testing.* New York: Springer.

## Concomitant Variable

It is not uncommon in designing research for an investigator to collect an array of variables representing characteristics on each observational unit. Some of these variables are central to the investigation, whereas others reflect preexisting differences in observational units and are not of interest per se. The latter are called *concomitant variables,* also referred to as *covariates.* Frequently in practice, these incidental variables represent undesired sources of variation influencing the dependent variable and are extraneous to the effects of manipulated (independent) variables, which are of primary interest. For designed experiments in which observational units are randomized to treatment conditions, failure to account for concomitant variables can exert a systematic influence (or bias) on the different treatment conditions. Alternatively, concomitant variables can increase the error variance, thereby reducing the likelihood of detecting real differences among the groups. Given these potential disadvantages, an ideal design strategy is one that would minimize the effect of the unwanted sources of variation corresponding to these concomitant variables. In practice, two general approaches are used to control the effects of concomitant variables: (1) experimental control and (2) statistical control. An expected benefit of controlling the effects of concomitant variables by either or both of these approaches is a substantial reduction in error variance, resulting in greater

precision for estimating the magnitude of treatment effects and increased statistical power.

## Experimental Control

Controlling the effects of concomitant variables is generally desirable. In addition to random assignment of subjects to experimental conditions, methods can be applied to control these variables in the design phase of a study. One approach is to use a small number of concomitant variables as the inclusion criteria for selecting subjects to participate in the study (e.g., only eighth graders whose parents have at least a high school education). A second approach is to match subjects on a small number of concomitant variables and then randomly assign each matched subject to one of the treatment conditions. This requires that the concomitant variables are available prior to the formation of the treatment groups. *Blocking,* or *stratification,* as it is sometimes referred to, is another method of controlling concomitant variables in the design stage of a study. The basic premise is that subjects are sorted into relatively homogeneous blocks on the basis of levels of one or two concomitant variables. The experimental conditions are subsequently randomized within each stratum. Exerting experimental control through case selection, matching, and blocking can reduce experimental error, often resulting in improved statistical power to detect differences among the treatment groups. As an exclusive design strategy, the usefulness of any one of these three methods to control the effects of concomitant variables is limited, however. It is necessary to recognize that countless covariates, in addition to those used to block or match subjects, may be affecting the dependent variable and thus posing potential threats to drawing appropriate inferences regarding treatment veracity. In contrast, randomization to experimental conditions ensures that any idiosyncratic differences among the groups are not systematic at the outset of the experiment. Random assignment does not guarantee that the groups are equivalent but rather that any observed differences are due only to chance.

## Statistical Control

The effects of concomitant variables can be controlled statistically if they are included in the models used in analyzing the data, for instance by the use of socioeconomic status as a covariate in an analysis of covariance (ANCOVA). Statistical control for regression procedures such as ANCOVA means removing from the experimental error and from the treatment effect all extraneous variance associated with the concomitant variable. This reduction in error variance is proportional to the strength of the linear relationship between the dependent variable and the covariate and is often quite substantial. Consequently, statistical control is most advantageous in situations in which the concomitant variable and outcome have a strong linear dependency (e.g., a covariate that represents an earlier administration of the same instrument used to measure the dependent variable).

In quasi-experimental designs in which random assignment of observational units to treatments is not possible or potentially unethical, statistical control is achieved through adjusting the estimated treatment effect by controlling for preexisting group differences on the covariate. This adjustment can be striking, especially when the difference on the concomitant variable across intact treatment groups is dramatic. Like blocking or matching, using ANCOVA to equate groups on important covariates should not be viewed as a substitute for randomization. Control of all potential concomitant variables is not possible in quasi-experimental designs, which therefore are always subject to threats to internal validity from unidentified covariates. This occurs because uncontrolled covariates may be confounded with the effects of the treatment in a manner such that group comparisons are biased.

## Benefits of Control

To the extent possible, research studies should be designed to control concomitant variables that are likely to systematically influence or mask the important relationships motivating the study. This can be accomplished by exerting experimental control through restricted selection procedures, randomization of subjects to experimental conditions, or stratification. In instances in which complete experimental control is not feasible or in conjunction with limited experimental control, statistical adjustments can be made through regression procedures such as ANCOVA. In either case,

advantages of instituting some level of control have positive benefits that include (a) a reduction in error variance and, hence, increased power and (b) a decrease in potential bias attributable to unaccounted concomitant variables.

*Jeffrey R. Harring*

***See also*** Analysis of Covariance (ANCOVA); Block Designs; Covariate; Experimental Design; Power; Statistical Control

**Further Readings**

Dayton, C. M. (1970). *The design of educational experiments*. New York: McGraw-Hill.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design and analysis: An integrated approach*. Hillsdale, NJ: Lawrence Erlbaum.

# Concurrent Validity

Validity refers to the degree to which a measure accurately taps the specific construct that it claims to be tapping. *Criterion-related validity* is concerned with the relationship between individuals' performance on two measures tapping the same construct. It typically is estimated by correlating scores on a new measure with scores from an accepted criterion measure. There are two forms of criterion-related validity: predictive validity and concurrent validity.

Concurrent validity focuses on the extent to which scores on a new measure are related to scores from a criterion measure administered at the same time, whereas *predictive validity* uses the scores from the new measure to predict performance on a criterion measure administered at a later time.

Examples of contexts in which concurrent validity is relevant include the following:

- Scores on a written first aid exam are highly correlated with scores assigned by raters during a hands-on measure in which examinees demonstrate first aid procedures.
- Scores on a depression inventory are used to classify individuals who are simultaneously diagnosed by a licensed psychologist.

The primary motives for developing a new measure designed to tap the same construct of interest as an established measure are cost and convenience. A new measure that is shorter or less expensive but leads users to draw the same conclusions as a longer, more costly measure is a highly desirable alternative. For example, to obtain the necessary information about examinee ability, decision makers might be able to administer a short written test to a large number of individuals at the same time instead of pricey, one-on-one performance evaluations that involve multiple ratings.

Before users make decisions based on scores from a new measure, they must have evidence that there is a close relationship between the scores of that measure and the performance on the criterion measure. This evidence can be obtained through a *concurrent validation study.* In a concurrent validation study, the new measure is administered to a sample of individuals that is representative of the group for which the measure will be used. An established criterion measure is administered to the sample at, or shortly after, the same time. The strength of the relationship between scores on the new measure and scores on the criterion measure indicates the degree of concurrent validity of the new measure.

The results of a concurrent validation study are typically evaluated in one of two ways, determined by the level of measurement of the scores from the two measures. When the scores on both the new measure and the criterion measure are *continuous,* the degree of concurrent validity is established via a *correlation coefficient,* usually the Pearson product–moment correlation coefficient. The correlation coefficient between the two sets of scores is also known as the *validity coefficient.* The validity coefficient can range from $-1$ to $+1$; coefficients close to 1 in absolute value indicate high concurrent validity of the new measure.

In the example shown in Figure 1, the concurrent validity of the written exam is quite satisfactory

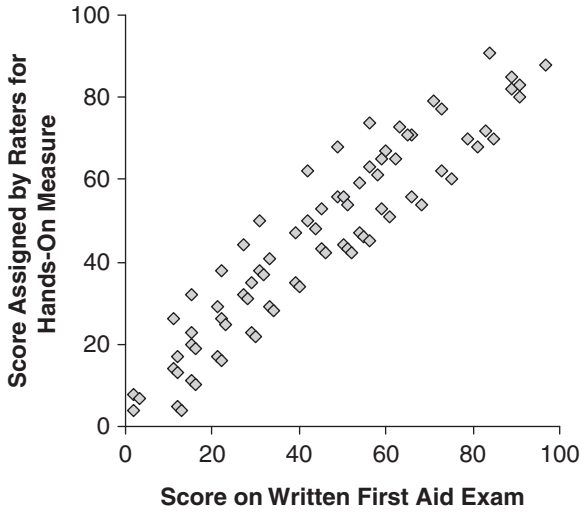



**Figure 1** Scatterplot of Scores on a Written First Aid Exam and Scores Assigned by Raters During a Hands-On Demonstration Measure

because the scores on the written exam correlate highly with the scores assigned by the rater; individuals scoring well on the written measure were also rated as highly proficient and vice versa.

When the new measure and the criterion measure are both used to classify individuals, coefficients of classification agreement, which are variations of correlation coefficients for categorical data, are typically used. Evidence of high concurrent validity is obtained when the classifications based on the new measure tend to agree with classifications based on the criterion measure.

Table 1 displays hypothetical results of a concurrent validation study for categorical data. The concurrent validity of the depression inventory is high because the classification based on the new measure and the psychologist's professional judgment match in 91% of the cases.

When determining the concurrent validity of a new measure, the selection of a valid criterion measure is critical. Robert M. Thorndike has noted that criterion measures should ideally be relevant to the desired decisions, free from bias, and reliable. In other words, they should already possess all the ideal measurement conditions that the new measure should possess also. Specifically, criterion measures should be

**Table 1** Cross-Classification of Classification From Inventory and Diagnosis by Psychologist

| *Classification from Depression Inventory* | *Diagnosis by Psychologist* | |
|---|---|---|
| | *Depressed* | *Not Depressed* |
| Depressed | 14 | 3 |
| Not Depressed | 7 | 86 |

- *relevant to the desired decisions:* Scores or classifications on the criterion measure should closely relate to, or represent, variation on the construct of interest. Previous validation studies and expert opinions should demonstrate the usefulness and appropriateness of the criterion for making inferences and decisions about the construct of interest.
- *free from bias:* Scores or classifications on the criterion measure should be free from bias, meaning that they should not be influenced by anything other than the construct of interest. Specifically, scores should not be impacted by personal characteristics of the individual, subjective opinions of a rater, or other measurement conditions.
- *reliable:* Scores or classifications on the criterion measure should be stable and replicable. That is, conclusions drawn about the construct of interest should not be clouded by inconsistent results across repeated administrations, alternative forms, or a lack of internal consistency of the criterion measure.

If the criterion measure against which the new measure is compared is invalid because it fails to meet these quality standards, the results of a concurrent validation study will be compromised. Put differently, the results from a concurrent validation study are only as useful as the quality of the criterion measure that is used in it. Thus, it is key to select the criterion measure properly to ensure that a lack of a relationship between the scores from the new measure and the scores from the criterion measure is truly due to problems with the new measure and not to problems with the criterion measure.

Of course, the selection of an appropriate criterion measure will also be influenced by the

availability or the cost of the measure. Thus, the practical limitations associated with criterion measurements that are inconvenient, expensive, or highly impractical to obtain may outweigh other desirable qualities of these measures.

*Jessica Lynn Mislevy and André A. Rupp*

***See also*** Criterion Validity; Predictive Validity

**Further Readings**

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Belmont, CA: Wadsworth Group.

Thorndike, R. M. (2005). *Measurement and evaluation in psychology and education* (7th ed.). Upper Saddle River, NJ: Pearson Education.

# Confidence Intervals

A confidence interval is an interval estimate of an unknown population parameter. It is constructed according to a random sample from the population and is always associated with a certain confidence level that is a probability, usually presented as a percentage. Commonly used confidence levels include 90%, 95%, and 99%. For instance, a confidence level of 95% indicates that 95% of the time the confidence intervals will contain the population parameter. A higher confidence level usually forces a confidence interval to be wider.

Confidence intervals have a long history. Using confidence intervals in statistical inference can be tracked back to the 1930s, and they are being used increasingly in research, especially in recent medical research articles. Researchersand research organizations such as the American Psychological Association suggest that confidence intervals should always be reported because confidence intervals provide information on both significance of test and variability of estimation.

## Interpretation

A confidence interval is a range in which an unknown population parameter is likely to be included. After independent samples are randomly selected from the same population, one confidence interval is constructed based on one sample with a certain confidence level. Together, all the confidence intervals should include the population parameter with the confidence level.

Suppose one is interested in estimating the proportion of bass among all types of fish in a lake. A 95% confidence interval for this proportion, [25%, 36%], is constructed on the basis of a random sample of fish in the lake. After more independent random samples of fish are selected from the lake, through the same procedure more confidence intervals are constructed. Together, all these confidence intervals will contain the true proportion of bass in the lake approximately 95% of the time.

The lower and upper boundaries of a confidence interval are called *lower confidence limit* and *upper confidence limit,* respectively. In the earlier example, 25% is the lower 95% confidence limit, and 36% is the upper 95% confidence limit.

## Confidence Interval Versus Significance Test

A significance test can be achieved by constructing confidence intervals. One can conclude whether a test is significant based on the confidence intervals. Suppose the null hypothesis is that the population mean $\mu$ equals 0 and the predetermined significance level is $\alpha$. Let $I$ be the constructed $100(1-\alpha)\%$ confidence interval for $\mu$. If 0 is included in the interval $I$, then the null hypothesis is accepted; otherwise, it is rejected. For example, a researcher wants to test whether the result of an experiment has a mean 0 at 5% significance level. After the 95% confidence interval $[-0.31, 0.21]$ is obtained, it can be concluded that the null hypothesis is accepted, because 0 is included in the confidence interval. If the 95% confidence interval is [0.11, 0.61], it indicates that the mean is significantly different from 0.

For a predetermined significance level or confidence level, the ways of constructing confidence intervals are usually not unique. Shorter confidence intervals are usually better because they indicate greater power in the sense of significance test.

One-sided significance tests can be achieved by constructing one-sided confidence intervals. Suppose a researcher is interested in an alternative hypothesis that the population mean is larger than

0 at the significance level .025. The researcher will construct a one-sided confidence interval taking the form of $(-\infty, b]$ with some constant $b$. Note that the width of a one-sided confidence interval is infinity. Following the above example, the null hypothesis would be that the mean is less than or equal to 0 at the .025 level. If the 97.5% one-sided confidence interval is $(-\infty, 3.8]$, then the null hypothesis is accepted because 0 is included in the interval. If the 97.5% confidence interval is $(-\infty, -2.1]$ instead, then the null hypothesis is rejected because there is no overlap between $(-\infty, -2.1]$ and $[0, \infty)$.

## Examples

### *Confidence Intervals for a Population Mean With Confidence Level* $100(1-\alpha)\%$

Confidence intervals for a population mean are constructed on the basis of the sample mean distribution.

#### The Population Follows a Normal Distribution With a Known Variance $\sigma^2$

After a random sample of size $N$ is selected from the population, one is able to calculate the sample mean $\overline{x}$, which is the average of all the observations. Thus the confidence interval is $[\overline{x} - z_{\alpha/2} \times \sigma/\sqrt{N}, \overline{x} + z_{\alpha/2} \times \sigma/\sqrt{N}]$ that is centered at $\overline{x}$ with half of the length $z_{\alpha/2} \times \sigma/\sqrt{N}$. $z_{\alpha/2}$ is the upper $\alpha/2$ quantile, meaning $P(z_{\alpha/2} \le Z) = \alpha/2$. Here $Z$ is a standard normal random variable. To find $z_{\alpha/2}$, one can either refer to the standard normal distribution table or use statistical software. Nowadays most of the statistical software, such as Excel, R, SAS, SPSS (an IBM company, formerly called PASW® Statistics), and Splus, have a simple command that will do the job. For commonly used confidence intervals, 90% confidence level corresponds to $z_{0.05} = 1.645$, 95% corresponds to $z_{0.025} = 1.96$, and 99% corresponds to $z_{0.005} = 2.56$.

The above confidence interval will shrink to a point if $N$ goes to infinity. So the interval estimate turns into a point estimate. It can be interpreted as if the whole population is taken as the sample; the sample mean is actually the population mean.

#### The Population Follows a Normal Distribution With Unknown Variance.

Suppose a sample of size $N$ is randomly selected from the population with observations $x_1, x_2, \ldots, x_N$. Let $\overline{x} = \sum_{i=1}^{N} x_i/N$. This is the sample mean. The sample variance is defined as $\sum_{i=1}^{N} (x_i - \overline{x})^2/(N-1)$, denoted by $s^2$. Therefore the confidence interval is $[\overline{x} - t_{N-1,\alpha/2} \times s/\sqrt{N}, \overline{x} + t_{N-1,\alpha/2} \times s/\sqrt{N}]$. Here $t_{N-1,\alpha/2}$ is the upper $\alpha/2$ quantile, meaning $P(t_{N-1,\alpha/2} \le T_{N-1}) = \alpha/2$ and $T_{N-1}$ follows the $t$ distribution with degree of freedom $N-1$. Refer to the $t$-distribution table or use software to get $t_{N-1,\alpha/2}$. If $N$ is greater than 30, one can use $z_{\alpha/2}$ instead of $t_{N-1,\alpha/2}$ in the confidence interval formula because there is little difference between them for large enough $N$.

As an example, the students' test scores in a class follow a normal distribution. One wants to construct a 95% confidence interval for the class average score based on an available random sample of size $N = 10$. The 10 scores are 69, 71, 77, 79, 82, 84, 80, 94, 78, and 67. The sample mean and the sample variance are 78.1 and 62.77, respectively. According the $t$-distribution table, $t_{9,0.25} = 2.26$. The 95% confidence interval for the class average score is [72.13, 84.07].

#### The Population Follows a General Distribution Other Than Normal Distribution

This is a common situation one might see in practice. After obtaining a random sample of size $N$ from the population, where it is required that $N \ge 30$, the sample mean and the sample variance can be computed as in the previous subsections, denoted by $\overline{x}$ and $s^2$, respectively. According to the central limit theorem, an approximate confidence interval can be expressed as $[\overline{x} - z_{\alpha/2} \times s/\sqrt{N}, \overline{x} + z_{\alpha/2} \times s/\sqrt{N}]$.

### *Confidence Interval for Difference of Two Population Means With Confidence Level* $100(1-\alpha)\%$

One will select a random sample from each population. Suppose the sample sizes are $N_1$ and $N_2$. Denote the sample means by $\overline{x}_1$ and $\overline{x}_2$

and the sample variances by $s_1^2$and $s_2^2$, respectively, for the two samples. The confidence interval is $[\overline{x}_1 - \overline{x}_2 - z_{\alpha/2} \times \sqrt{s_1^2/N_1 + s_2^2/N_2}$, $\overline{x}_1 - \overline{x}_2 + z_{\alpha/2} \times \sqrt{s_1^2/N_1 + s_2^2/N_2}]$.

If one believes that the two population variances are about the same, the confidence interval will be $[\overline{x}_1 - \overline{x}_2 - t_{N_1+N_2-2,\alpha/2} \times s_p, \overline{x}_1 - \overline{x}_2 + t_{N_1+N_2-2,\alpha/2} \times s_p]$, where $s_p = \sqrt{(N_1+N_2)[(N_1-1)s_1^2 + (N_2-1)s_2^2]/[(N_1+N_2-2)N_1N_2]}$.

Continuing with the above example about the students' scores, call that class Class A: If one is interested in comparing the average scores between Class A and another class, Class B, the confidence interval for the difference between the average class scores will be constructed. First, randomly select a group of students in Class B. Suppose the group size is 8. These eight students' scores are 68, 79, 59, 76, 80, 89, 67, and 74. The sample mean is 74, and the sample variance is 86.56 for Class B. If Class A and Class B are believed to have different population variances, then the 95% confidence interval for the difference of the average scores is [−4.03, 12.17] by the first formula provided in this subsection. If one believes these two classes have about the same population variance, then the 95% confidence interval will be changed to [−4.47, 12.57] by the second formula.

### *Confidence Intervals for a Single Proportion and Difference of Two Proportions*

Sometimes one may need to construct confidence intervals for a single unknown population proportion. Denote $\hat{p}$ the sample proportion, which can be obtained from a random sample from the population. The estimated standard error for the proportion is $se(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})/N}$. Thus the confidence interval for the unknown proportion is $[\hat{p} - z_{\alpha/2} \times se(\hat{p}), \hat{p} + z_{\alpha/2} \times se(\hat{p})]$.

This confidence interval is constructed on the basis of the normal approximation (refer to the central limit theorem for the normal approximation). The normal approximation is not appropriate when the proportion is very close to 0 or 1. A rule of thumb is that when $N\hat{p} > 5$ and $N(1-\hat{p}) > 5$, usually the normal approximation works well.

For example, a doctor wants to construct a 99% confidence interval for the chance of having a certain disease by studying patients' x-ray slides. $N = 30$ x-ray slides are randomly selected, and the number of positive slides follows a distribution known as the binomial distribution. Suppose 12 of them are positive for the disease. Hence $\hat{p} = 12/30$, which is the sample proportion. Since $N\hat{p} = 12$ and $N(1-\hat{p}) = 18$ are both larger than 5, the confidence interval for the unknown proportion can be constructed using the normal approximation. The estimated standard error for $\hat{p}$ is 0.09. Thus the lower 99% confidence limit is 0.17 and the upper 99% confidence limit is 0.63. So the 99% confidence interval is $[0.17, 0.63]$.

The range of a proportion is between 0 and 1. But sometimes the constructed confidence interval for the proportion may exceed it. When this happens, one should truncate the confidence interval to make the lower confidence limit 0 or the upper confidence limit 1.

Since the binomial distribution is discrete, a correction for continuity of $0.5/N$ may be used to improve the performance of confidence intervals. The corrected upper limit is added by $0.5/N$, and the corrected lower limit is subtracted by $0.5/N$.

One can also construct confidence intervals for the proportion difference between two populations based on the normal approximation. Suppose two random samples are independently selected from the two populations, with sample sizes $N_1$ and $N_2$ and sample proportions $\hat{p}_1$ and $\hat{p}_2$ respectively. The estimated population proportion difference is the sample proportion difference, $\hat{p}_1 - \hat{p}_2$, and the estimated standard error for the proportion difference is $se(\hat{p}_1 - \hat{p}_2) = \sqrt{\hat{p}_1(1-\hat{p}_1)/N_1 + \hat{p}_2(1-\hat{p}_2)/N_2}$. The confidence interval for two-sample proportion difference is $[(\hat{p}_1 - \hat{p}_2) - z_{\alpha/2} \times se(\hat{p}_1 - \hat{p}_2), (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2} \times se(\hat{p}_1 - \hat{p}_2)]$.

Similar to the normal approximation for a single proportion, the approximation for the proportion difference depends on sample sizes and sample proportions. The rule of thumb is that $N\hat{p}_1$, $N(1-\hat{p}_1)$, $N\hat{p}_2$, and $N(1-\hat{p}_2)$ should each be larger than 10.

### *Confidence Intervals for Odds Ratio*

An *odds ratio* (OR) is a commonly used effect size for categorical outcomes, especially in health

science, and is the ratio of the odds in Category 1 to the odds in Category 2. For example, one wants to find out the relationship between smoking and lung cancer. Two groups of subjects, smokers and nonsmokers, are recruited. After a few years' follow-ups, $N_{11}$ subjects among the smokers are diagnosed with lung cancer and $N_{21}$ subjects among the nonsmokers. There are $N_{12}$ and $N_{22}$ subjects who do not have lung cancer among the smokers and the nonsmokers, respectively. The *odds* of having lung cancer among the smokers and the nonsmokers are estimated as $N_{11}/N_{12}$ and $N_{21}/N_{22}$, respectively. The OR of having lung cancer among the smokers compared with the nonsmokers is the ratio of the above two odds, which is

$$(N_{11}/N_{12})/(N_{21}/N_{22}) = (N_{11}N_{22})/(N_{21}N_{12}).$$

For a relatively large total sample size, $\ln(OR)$ is approximated to a normal distribution, so the construction for the confidence interval for $\ln(OR)$ is similar to that for the normal distribution. The standard error for $\ln(OR)$ is defined as $se(\ln(OR)) = \sqrt{1/N_{11} + 1/N_{12} + 1/N_{21} + 1/N_{22}}$. A 95% confidence interval for $\ln(OR)$ is $[\ln(OR) - 1.96 \times se(\ln(OR)), \ln(OR) + 1.96 \times se(\ln(OR))]$. As the exponential function is monotonic, there is one-to-one mapping between the OR and $\ln(OR)$. Thus a 95% confidence interval for the OR is $[\exp(\ln(OR) - 1.96 \times se(\ln(OR))), \exp(\ln(OR) + 1.96 \times se(\ln(OR)))]$.

The confidence intervals for the OR are not symmetric about the estimated $OR$. But one can still tell the significance of the test on the basis of the corresponding confidence interval for the OR. For the above example, the null hypothesis is that there is no difference between smokers and nonsmokers in the development of lung cancer; that is, $OR = 1$. If 1 is included in the confidence interval, one should accept the null hypothesis; otherwise, one should reject it.

### *Confidence Intervals for Relative Risk*

Another widely used concept in health care is *relative risk* (RR), which is the risk difference between two groups. *Risk* is defined as the chance of having a specific outcome among subjects in that group. Taking the above example, the risk of having lung cancer among smokers is estimated as $N_{11}/(N_{11} + N_{12})$, and the risk among nonsmokers is estimated as $N_{21}/(N_{21} + N_{22})$. The RR is the ratio of the above two risks, which is $[N_{11}/(N_{11} + N_{12})]/[N_{21}/(N_{21} + N_{22})]$.

**Table 1** Lung Cancer Among Smokers and Nonsmokers

| *Smoking Status* | *Lung Cancer* | *No Lung Cancer* | *Total* |
|---|---|---|---|
| Smokers | $N_{11}$ | $N_{12}$ | $N_{1.}$ |
| Nonsmokers | $N_{21}$ | $N_{22}$ | $N_{2.}$ |
| Total | $N_{.1}$ | $N_{.2}$ | $N$ |

Like the OR, the sampling distribution of $\ln(RR)$ is approximated to a normal distribution. The standard error for $\ln(RR)$ is $se(\ln(RR)) = \sqrt{1/N_{11} - 1/N_{1.} + 1/N_{21} - 1/N_{2.}}$. A 95% confidence interval for $\ln(RR)$ is

$$[\ln(RR) - 1.96 \times se(\ln(RR)), \ln(RR) + 1.96 \times se(\ln(RR))].$$

Thus the 95% confidence interval for the RR is $[\exp(\ln(RR) - 1.96 \times se(\ln(RR))), \exp(\ln(RR) + 1.96 \times se(\ln(RR)))]$.

The confidence intervals for the RRs are not symmetric about the estimated $RR$ either. One can tell the significance of a test from the corresponding confidence interval for the RR. Usually the null hypothesis is that $RR = 1$, which means that the two groups have the same risk. For the above example, the null hypothesis would be that the risks of developing lung cancer among smokers and nonsmokers are equal. If 1 is included in the confidence interval, one may accept the null hypothesis. If not, the null hypothesis should be rejected.

### *Confidence Intervals for Variance*

A confidence interval for unknown population variance can be constructed with the use of a central chi-square distribution. For a random sample with size $N$ and sample variance $s^2$, an approximate two-sided $100(1-\alpha)\%$ confidence interval for population variance $\sigma^2$ is $[(N-1)s^2/\chi^2_{N-1}(\alpha/2), (N-1)s^2/\chi^2_{N-1}(1-\alpha/2)]$. Here $\chi^2_{N-1}(\alpha/2)$ is at the upper $\alpha/2$ quantile satisfying the requirement that the probability that a central chi-square random variable with degree of freedom $N-1$ is greater than $\chi^2_{N-1}(\alpha/2)$ is $\alpha/2$. Note that this confidence interval may not work well if the sample size is small or the distribution is far from normal.

## Bootstrap Confidence Intervals

The bootstrap method provides an alternative way for constructing an interval to measure the accuracy of an estimate. It is especially useful when the usual confidence interval is hard or impossible to calculate. Suppose $s(x)$ is used to estimate an unknown population parameter $\theta$ based on a sample $x$ of size $N$. A bootstrap confidence interval for the estimate can be constructed as follows. One randomly draws another sample $x^*$ of the same size $N$ with replacement from the original sample $x$. The estimate $s(x^*)$ based on $x^*$ is called a *bootstrap replication* of the original estimate. Repeat the procedure for a large number of times, say 1,000 times. Then the $\alpha$th quantile and the $(1-\alpha)$th quantile of the 1,000 bootstrap replications serve as the lower and upper limits of a $100(1-2\alpha)\%$ bootstrap confidence interval. Note that the interval obtained in this way may vary a little from time to time due to the randomness of bootstrap replications. Unlike the usual confidence interval, the bootstrap confidence interval does not require assumptions on the population distribution. Instead, it highly depends on the data $x$ itself.

## Simultaneous Confidence Intervals

Simultaneous confidence intervals are intervals for estimating two or more parameters at a time. For example, suppose $\mu_1$ and $\mu_2$ are the means of two different populations. One wants to find confidence intervals $I_1$ and $I_2$ simultaneously such that

$$P(\mu_1 \in I_1 \text{ and } \mu_2 \in I_2) = 1-\alpha.$$

If the sample $x_1$ used to estimate $\mu_1$ is independent of the sample $x_2$ for $\mu_2$, then $I_1$ and $I_2$ can be simply calculated as $100\sqrt{1-\alpha}\%$ confidence intervals for $\mu_1$ and $\mu_2$, respectively.

The simultaneous confidence intervals $I_1$ and $I_2$ can be used to test whether $\mu_1$ and $\mu_2$ are equal. If $I_1$ and $I_2$ are nonoverlapped, then $\mu_1$ and $\mu_2$ are significantly different from each other at a level less than $\alpha$.

Simultaneous confidence intervals can be generalized into a *confidence region* in the multidimensional parameter space, especially when the estimates for parameters are not independent. A $100(1-\alpha)\%$ confidence region $D$ for the parameter vector $\theta$ satisfies $P(\theta \in D) = 1-\alpha$, where $D$ does not need to be a product of simple intervals.

## Links With Bayesian Intervals

A Bayesian interval, or *credible interval,* is derived from the posterior distribution of a population parameter. From a Bayesian point of view, the parameter $\theta$ can be regarded as a random quantity, which follows a distribution $P(\theta)$, known as the *prior distribution.* For each fixed $\theta$, the data $x$ are assumed to follow the conditional distribution $P(x|\theta)$, known as the *model.* Then Bayes's theorem can be applied on $x$ to get the adjusted distribution $P(\theta|x)$ of $\theta$, known as the *posterior distribution,* which is proportional to $P(x|\theta) \cdot P(\theta)$. A $1-\alpha$ Bayesian interval $I$ is an interval that satisfies $P(\theta \in I|x) = 1-\alpha$ according to the posterior distribution. In order to guarantee the optimality or uniqueness of $I$, one may require that the values of the posterior density function inside $I$ always be greater than any one outside $I$. In those cases, the Bayesian interval $I$ is called the *highest posterior density region.* Unlike the usual confidence intervals, the level $1-\alpha$ of a Bayesian interval $I$ indicates the probability that the random $\theta$ falls into $I$.

*Qiaoyan Hu, Shi Zhao, and Jie Yang*

***See also*** Bayes's Theorem; Boostrapping; Central Limit Theorem; Normal Distribution; Odds Ratio; Sample; Significance, Statistical

### Further Readings

Blyth, C. R., & Still, H. A. (1983). Binomial confidence intervals. *Journal of the American Statistical Association, 78,* 108–116.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall/CRC.

Fleiss, J. R., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions*. Hoboken, NJ: Wiley.

Newcombe, R. G. (1998). Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Statistics in Medicine, 17,* 857–872.

Pagano, M., & Gauvreau, K. (2000). *Principles of biostatistics*. Pacific Grove, CA: Duxbury.

Smithson, M. (2003).*Confidence intervals*. Thousand Oaks, CA: Sage.

Stuart, A., Ord, K., & Arnold, S. (1998). *Kendall's advanced theory of statistics: Vol. 2A. Classical inference and the linear model* (6th ed.). London: Arnold.

# Confirmatory Factor Analysis

Research in the social and behavioral sciences often focuses on the measurement of unobservable, theoretical constructs such as ability, anxiety, depression, intelligence, and motivation. Constructs are identified by directly observable, manifest variables generally referred to as indicator variables (note that *indicator, observed,* and *manifest variables* are often used interchangeably). Indicator variables can take many forms, including individual items or one or more composite scores constructed across multiple items. Many indicators are available for measuring a construct, and each may differ in how reliably it measures the construct. The choice of which indicator to use is based largely on availability. Traditional statistical techniques using single indicator measurement, such as regression analysis and path analysis, assume the indicator variable to be an error-free measure of the particular construct of interest. Such an assumption can lead to erroneous results.

Measurement error can be accounted for by the use of multiple indicators of each construct, thus creating a *latent variable.* This process is generally conducted in the framework of factor analysis, a multivariate statistical technique developed in the early to mid-1900s primarily for identifying and/or validating theoretical constructs. The general factor analysis model can be implemented in an exploratory or confirmatory framework. The exploratory framework, referred to as *exploratory factor analysis* (EFA), seeks to explain the relationships among the indicator variables through a given number of previously undefined latent variables. In contrast, the confirmatory framework, referred to as *confirmatory factor analysis* (CFA), uses latent variables to reproduce and test previously defined relationships between the indicator variables. The methods differ in their underlying purpose. Whereas EFA is a data-driven approach, CFA is a hypothesis driven approach requiring theoretically and/or empirically based insight into the relationships among the indicator variables. This insight is essential for establishing a starting point for the specification of a model to be tested. What follows is a more detailed, theoretical discussion of the CFA model, the process of implementing the CFA model, and the manner and settings in which the CFA model is most commonly implemented.

## The Confirmatory Factor Analysis Model

The CFA model belongs to the larger family of modeling techniques referred to as *structural equation models* (SEM). Structural equation models offer many advantages over traditional modeling techniques (although many traditional techniques, such as multiple regression, are considered special types of SEMs), among these the use of latent variables and the ability to model complex measurement error structures. The latter, measurement error, is not accounted for in the traditional techniques, and, as was previously mentioned, ignoring measurement error oftentimes leads to inaccurate results. Many different types of models can be put forth in the SEM framework, with the more elaborate models containing both of the SEM submodels: (a) the measurement model and (b) the structural model. The measurement model uses latent variables to explain variability that is shared by the indicator variables and variability that is unique to each indicator variable. The structural model, on the other hand, builds on the measurement model by analyzing the associations between the latent variables as direct causal effects.

The CFA model is considered a measurement model in which an unanalyzed association (covariance) between the latent variables is assumed to exist. Figure 1 displays a basic two-factor CFA model in LISREL notation. This diagram displays two latent variables, labeled with xi (pronounced ksi), $\xi_1$ and $\xi_2$, each identified by three indicator variables (i.e., $X_1$, $X_2$, and $X_3$ for $\xi_1$ and $X_4$, $X_5$, and $X_6$ for $\xi_2$). The single-headed arrows pointing from the latent variables to each of the indicator variables represent factor loadings or pattern coefficients, generally interpreted as regression coefficients, labeled with lambda as $\lambda_{11}, \lambda_{21}, \lambda_{31}, \lambda_{42}, \lambda_{52}$, and $\lambda_{62}$ and signifying the direction and magnitude of the relationship between each indicator and latent variable. For instance, $\lambda_{21}$ represents the direction and magnitude of the relationship between latent variable one ($\xi_1$) and indicator variable two ($X_2$). These coefficients can be reported in standardized or unstandardized form. The double-headed arrow between the latent variables,

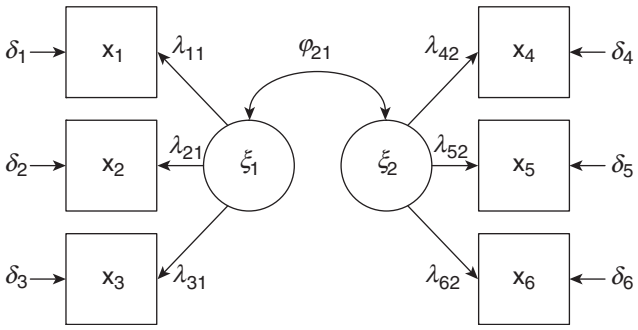



**Figure 1** Two-Factor Confirmatory Factor Analysis in LISREL Notation

labeled as $\varphi_{21}$ (phi), represents an unanalyzed covariance or correlation between the latent variables. This value signifies the direction and magnitude of the relationship between the latent variables. Variability in the indicator variables not attributable to the latent variable, termed measurement error, is represented by the single-headed arrows pointing toward the indicator variables from each of the measurement error terms, labeled with delta as $\delta_1, \delta_2, \delta_3, \delta_4, \delta_5,$ and $\delta_6$. These are essentially latent variables assumed to be independent of each other and of the latent variables $\xi_1$ and $\xi_2$.

The CFA model depicted in Figure 1 can be represented in equation form as follows:

$$
\begin{aligned}
x1 &= \lambda 11\xi 1 + \delta 1 \\
x2 &= \lambda 21\xi 1 + \delta 2 \\
x3 &= \lambda 31\xi 1 + \delta 3 \\
x4 &= \lambda 42\xi 2 + \delta 4 \\
x5 &= \lambda 52\xi 2 + \delta 5 \\
x6 &= \lambda 62\xi 2 + \delta 6
\end{aligned}
$$

These equations can be written in more compact form by using matrix notation:

$$x = \Lambda\xi + \delta,$$

where $x$ is a $q \times 1$ vector of observed indicator variables (where $q$ is the number of observed indicator variables), $\Lambda$ is a $q \times n$ matrix of regression coefficients relating the indicators to the latent variables (where $n$ is the number of latent variables), $\xi$ is an $n \times 1$ vector of latent variables, and $\delta$ is a $q \times 1$ vector of measurement errors that are assumed to be independent of each other. The equation above written in matrix notation is recognized as the measurement model in a general structural equation model.

The path diagram in Figure 1 depicts an ideal CFA model. The terms *simple structure* and *unidimensional measurement* are used in the general factor analysis framework to signify a model that meets two conditions: (1) Each latent variable is defined by a subset of indicator variables which are strong indicators of that latent variable, and (2) each indicator variable is strongly related to a single latent variable and weakly related to the other latent variable(s). In general, the second condition is met in the CFA framework with the initial specification of the model relating each indicator variable to a single latent variable. However, subsequent testing of the initial CFA model may reveal the presence of an indicator variable that is strongly related to multiple latent variables. This is a form of multidimensional measurement represented in the model by a path between an indicator variable and the particular latent variable it is related to. This is referred to as a *cross-loading* (not included in Figure 1).

Researchers using EFA often encounter cross-loadings because models are not specified in advance. Simple structure can be obtained in EFA through the implementation of a rotation method. However, a rotated factor solution does not remove the presence of cross-loadings, particularly when the rotation method allows for correlated factors (i.e., oblique rotation). As previously mentioned, specifying the model in advance provides the researcher flexibility in testing if the inclusion of one or more cross-loadings results in a better explanation of the relationships in the observed data. Literature on this topic reveals the presence of conflicting points of view. In certain instances, it may be plausible for an indicator to load on multiple factors. However, specifying unidimensional models has many advantages, including but not limited to providing a clearer interpretation of the model and a more precise assessment of convergent and discriminant validity than multidimensional models do.

Multidimensional measurement is not limited to situations involving cross-loadings. Models are also considered multidimensional when measurement

errors are not independent of one another. In such models, measurement errors can be specified to correlate. Correlating measurement errors allows for hypotheses to be tested regarding shared variance that is not due to the underlying factors. Specifying the presence of correlated measurement errors in a CFA model should be based primarily on model parsimony and, perhaps most important, on substantive considerations.

## Conducting a Confirmatory Factor Analysis

Structural equation modeling is also referred to as *covariance structure analysis* because the covariance matrix is the focus of the analysis. The general null hypothesis to be tested in CFA is

$$\Sigma = \Sigma(\Theta),$$

where $\Sigma$ is the population covariance matrix, estimated by the sample covariance matrix $S$, and $\Sigma(\theta)$ is the model-implied covariance matrix, estimated by $\Sigma(\Theta)$. Researchers using CFA seek to specify a model that most precisely explains the relationships among the variables in the original data set. In other words, the model put forth by the researcher reproduces or fits the observed sample data to some degree. The more precise the fit of the model to the data, the smaller the difference between the sample covariance matrix, $S$, and the model-implied covariance matrix, $\Sigma(\Theta)$. To evaluate the null hypothesis in this manner, a sequence of steps common to implementing structural equation models must be followed.

The first step, often considered the most challenging, requires the researcher to specify the model to be evaluated. To do so, researchers must use all available information (e.g., theory, previous research) to postulate the relationships they expect to find in the observed data prior to the data collection process. Postulating a model may also involve the use of EFA as an initial step in developing the model. Different data sets must be used when this approach is taken because EFA results are subject to capitalizing on chance. Following an EFA with a CFA on the same data set may compound the capitalization-on-chance problem and lead to inaccurate conclusions based on the results.

Next, the researcher must determine whether the model is identified. At this stage, the research must determine whether, based on the sample covariance matrix, $S$, and the model-implied covariance matrix, $\Sigma(\Theta)$, a unique estimate of each unknown parameter in the model can be identified. The model is identified if the number of parameters to be estimated is less than or equal to the number of unique elements in the variance–covariance matrix used in the analysis. This is the case when the degrees of freedom for the model are greater than or equal to 0. In addition, a metric must be defined for every latent variable, including the measurement errors. This is typically done by setting the metric of each latent variable equal to the metric of one of its indicators (i.e., fixing the loading between indicator and its respective latent variable to 1) but can also be done by setting the variance of each latent variable equal to 1.

The process of model estimation in CFA (and SEM, in general) involves the use of a fitting function such as generalized least squares or maximum likelihood (ML) to obtain estimates of model parameters that minimize the discrepancy between the sample covariance matrix, $S$, and the model implied covariance matrix, $\Sigma(\Theta)$. ML estimation is the most commonly used method of model estimation. All the major software packages (e.g., AMOS, EQS, LISREL, MPlus) available for posing and testing CFA models provide a form of ML estimation. This method has also been extended to address issues that are common in applied settings (e.g., nonnormal data, missing data ), making CFA applicable to a wide variety of data types.

Various goodness-of-fit indices are available for determining whether the sample covariance matrix, $S$, and the model implied covariance matrix, $\Sigma(\Theta)$, are sufficiently equal to deem the model meaningful. Many of these indices are derived directly from the ML fitting function. The primary index is the chi-square. Because of the problems inherent with this index (e.g., inflated by sample size), assessing the fit of a model requires the use of multiple indices from the three broadly defined categories of indices: (1) absolute fit, (2) parsimony correction, and (3) comparative fit. The extensive research on fit indices has fueled the debate and answered many questions as to which are useful and what cutoff values should be adopted for determining adequate model fit in a variety of situations. Research has indicated that fit indices from each of these categories provide

pertinent information and should be used in unison when conducting a CFA.

Even for a well-fitting model, further respecification may be required because goodness-of-fit indices evaluate the model globally. Obviously, a model that fits the data poorly requires reconsideration of the nature of the relationships specified in the model. Models that fit the data at an acceptable level, however, may fail to explain all the relationships in the variance–covariance matrix adequately. The process of model respecification involves a careful examination of areas in which the model may not explain relationships adequately. A careful inspection of the residual variance–covariance matrix, which represents the difference between the sample variance–covariance matrix, $S$, and the model-implied variance–covariance matrix, $\Sigma(\Theta)$, can reveal poorly explained relationships. Modification indices can also be used for determining what part of a model should be respecified. The *Lagrange multiplier test* can be used to test whether certain relationships should be included in the model, and the *Wald test* can be used to determine whether certain relationships should be excluded from the model. Last, parameter estimates themselves should be carefully examined in terms of practical (i.e., interpretability) as well as statistical significance.

Following the aforementioned steps is common practice in the application of CFA models in research. The final step, respecification of the model, can uncover statistically equivalent versions of the original CFA model found to provide the best fit. These models are equivalent in the sense that they each explain the original data equally well. While many of the equivalent models that can be generated will not offer theoretically plausible alternatives to the original model, there will likely be a competing model that must be considered before a final model is chosen. The complex decision of which model offers the best fit statistically and theoretically can be made easier by a researcher with a strong grasp of the theory underlying the original model put forth.

## Applications of Confirmatory Factor Analysis Models

CFA has many applications in the practical realm. The following section discusses some of the more common uses of the CFA model.

### *Scale Development*

CFA can be particularly useful in the process of developing a scale for tests or survey instruments. Developing such an instrument requires a strong understanding of the theoretical constructs that are to be evaluated by the instrument. CFA procedures can be used to verify the underlying structure of the instrument and to quantify the relationship between each item and the construct it was designed to measure, as well as quantify the relationships between the constructs themselves. In addition, CFA can be used to aid in the process of devising a scoring system as well as evaluating the reliability of an instrument. Recent advances in the use of CFA with categorical data also make it an attractive alternative to the item response theory model for investigating issues such as measurement invariance.

### *Validity of Theoretical Constructs*

CFA is a particularly effective tool for examining the validity of theoretical constructs. One of the most common approaches to testing validity of theoretical constructs is through the multitrait–multimethod (MTMM) design, in which multiple traits (i.e., constructs) are assessed through multiple methods. The MTMM design is particularly effective in evaluating *convergent validity* (i.e., strong relations among different indicators, collected via different methods, of the same construct) and *discriminant validity* (i.e., weak relations among different indicators, collected via different methods, of distinct constructs). The MTMM design can also be used to determine whether a *method effect* (i.e., strong relations among different indicators of distinct constructs, collected via the same method) is present.

### *Measurement Invariance*

CFA models are commonly used in the assessment of measurement invariance and population heterogeneity. By definition, *measurement invariance* refers to the degree to which measurement models generalize across different groups (e.g., males vs. females, native vs. nonnative English speakers) as well as how a measurement instrument generalizes across time. To assess measurement invariance, CFA models can be specified as

either *multiple group models* or *multiple indicator–multiple cause (MIMIC) models*. While these models are considered interchangeable, there are advantages to employing the multiple group approach. In the multiple group approach, tests of invariance across a greater number of parameters can be conducted. The MIMIC model tests for differences in intercept and factor means, essentially providing information about which covariates have direct effects in order to determine what grouping variables might be important in a multiple group analysis. A multiple group model, on the other hand, offers tests for differences in intercept and factor means as well as tests of other parameters such as the factor loadings, error variances and covariances, factor means, and factor covariances. The caveat of employing a multiple group model, however, lies in the necessity of having a sufficiently large sample size for each group, as well as addressing the difficulties that arise in analyses involving many groups. The MIMIC model is a more practical approach with smaller samples.

### *Higher Order Models*

The CFA models presented to this point have been first-order models. These first-order models include the specification of all necessary parameters excluding the assumed relationship between the factors themselves. This suggests that even though a relationship between the factors is assumed to exist, the nature of that relationship is "unanalyzed" or not specified in the initial model. Higher order models are used in cases in which the relationship between the factors is of interest. A higher order model focuses on examining the relationship between the first-order factors, resulting in a distinction between variability shared by the first-order factors and variability left unexplained. The process of conducting a CFA with second-order factors is essentially the same as the process of testing a CFA with first-order factors.

### Summary

CFA models are used in a variety of contexts. Their popularity results from the need in applied research for formal tests of theories involving unobservable latent constructs. In general, the popularity of SEM and its use in testing causal relationships among constructs will require sound measurement of the latent constructs through the CFA approach.

*Greg William Welch*

***See also*** Exploratory Factor Analysis; Latent Variable; Structural Equation Modeling

#### Further Readings

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford Press.

Spearman, C. E. (1904). General intelligence objectively determined and measured. *American Journal of Psychology, 5,* 201–293.

Thurstone, L. L. (1935). *The vectors of the mind*. Chicago: University of Chicago Press.

## Confounding

Confounding occurs when two variables systematically covary. Researchers are often interested in examining whether there is a relationship between two or more variables. Understanding the relationship between or among variables, including whether those relationships are causal, can be complicated when an independent or predictor variable covaries with a variable other than the dependent variable. When a variable systematically varies with the independent variable, the confounding variable provides an explanation other than the independent variable for changes in the dependent variable.

### Confounds in Correlational Designs

Confounding variables are at the heart of the third-variable problem in correlational studies. In a correlational study, researchers examine the relationship between two variables. Even if two variables are correlated, it is possible that a third, confounding variable is responsible for the apparent relationship between the two variables. For example, if there were a correlation between icecream consumption and homicide rates, it would be a mistake to assume that eating ice cream causes homicidal rages or that murderers seek frozen treats after killing. Instead, a third

variable—heat—is likely responsible for both increases in ice cream consumption and homicides (given that heat has been shown to increase aggression). Although one can attempt to identify and statistically control for confounding variables in correlational studies, it is always possible that an unidentified confound is producing the correlation.

## Confounds in Quasi-Experimental and Experimental Designs

The goal of quasi-experimental and experimental studies is to examine the effect of some treatment on an outcome variable. When the treatment systematically varies with some other variable, the variables are confounded, meaning that the treatment effect is comingled with the effects of other variables. Common sources of confounding include history, maturation, instrumentation, and participant selection. History confounds may arise in quasi-experimental designs when an event that affects the outcome variable happens between pretreatment measurement of the outcome variable and its posttreatment measurement. The events that occur between pre- and posttest measurement, rather than the treatment, may be responsible for changes in the dependent variable. Maturation confounds are a concern if participants could have developed—cognitively, physically, emotionally—between pre- and posttest measurement of the outcome variable. Instrumentation confounds occur when different instruments are used to measure the dependent variable at pre- and posttest or when the instrument used to collect the observation deteriorates (e.g., a spring loosens or wears out on a key used for responding in a timed task). Selection confounds may be present if the participants are not randomly assigned to treatments (e.g., use of intact groups, participants self-select into treatment groups). In each case, the confound provides an alternative explanation—an event, participant development, instrumentation changes, preexisting differences between groups—for any treatment effects on the outcome variable.

Even though the point of conducting an experiment is to control the effects of potentially confounding variables through the manipulation of an independent variable and random assignment of participants to experimental conditions, it is possible for experiments to contain confounds. An experiment may contain a confound because the experimenter intentionally or unintentionally manipulated two constructs in a way that caused their systematic variation. The Illinois Pilot Program on Sequential, Double-Blind Procedures provides an example of an experiment that suffers from a confound. In this study commissioned by the Illinois legislature, eyewitness identification procedures conducted in several Illinois police departments were randomly assigned to one of two conditions. For the sequential, double-blind condition, administrators who were blind to the suspect's identity showed members of a lineup to an eyewitness sequentially (i.e., one lineup member at a time). For the single-blind, simultaneous condition, administrators knew which lineup member was the suspect and presented the witness with all the lineup members at the same time. Researchers then examined whether witnesses identified the suspect or a known-innocent lineup member at different rates depending on the procedure used. Because the mode of lineup presentation (simultaneous vs. sequential) and the administrator's knowledge of the suspect's identity were confounded, it is impossible to determine whether the increase in suspect identifications found for the single-blind, simultaneous presentations is due to administrator knowledge, the mode of presentation, or some interaction of the two variables. Thus, manipulation of an independent variable protects against confounding only when the manipulation cleanly varies a single construct.

Confounding can also occur in experiments if there is a breakdown in the random assignment of participants to conditions. In applied research, it is not uncommon for partners in the research process to want an intervention delivered to people who deserve or are in need of the intervention, resulting in the funneling of different types of participants into the treatment and control conditions. Random assignment can also fail if the study's sample size is relatively small because in those situations even random assignment may result in people with particular characteristics appearing in treatment conditions rather than in control conditions merely by chance.

### Statistical Methods for Dealing With Confounds

When random assignment to experimental conditions is not possible or is attempted but fails, it is likely that people in the different conditions also differ on other dimensions, such as attitudes, personality traits, and past experience. If it is possible to collect data to measure these confounding variables, then statistical techniques can be used to adjust for their effects on the causal relationship between the independent and dependent variables. One method for estimating the effects of the confounding variables is the calculation of *propensity scores*. A propensity score is the probability of receiving a particular experimental treatment condition given the participants' observed score on a set of confounding variables. Controlling for this propensity score provides an estimate of the true treatment effect adjusted for the confounding variables. The propensity score technique cannot control for the effects of unmeasured confounding variables. Given that it is usually easy to argue for additional confounds in the absence of clean manipulations of the independent variable and random assignment, careful experimental design that rules out alternative explanations for the effects of the independent variable is the best method for eliminating problems associated with confounding.

*Margaret Bull Kovera*

***See also*** Experimental Design; Instrumentation; Propensity Score Analysis; Quasi-Experimental Designs; Selection

#### Further Readings

McDermott, K. B., & Miller, G. E. (2007). Designing studies to avoid confounds. In R. S. Sternberg, H. L. Roediger, III, & D. F. Halpern (Eds.), *Critical thinking in psychology* (pp. 131–142). New York: Cambridge University Press.

Schacter, D. L., Daws, R., Jacoby, L. L., Kahneman, D., Lempert, R., Roediger, H. L., et al. (2008). Studying eyewitness investigations in the field. *Law & Human Behavior, 32,* 3–5.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin.

Vanderweele, T. (2006). The use of propensity score methods in psychiatric research. *International Journal of Methods in Psychiatric Research, 15,* 95–103.

## Congruence

The *congruence* between two configurations of points quantifies their similarity. The configurations to be compared are, in general, produced by factor analytic methods that decompose an "observations by variables" data matrix and produce one set of factor scores for the observations and one set of factor scores (i.e., the *loadings*) for the variables. The congruence between two sets of factor scores collected on the same units (which can be observations or variables) measures the similarity between these two sets of scores. If, for example, two different types of factor analysis are performed on the same data set, the *congruence* between the two solutions is evaluated by the similarity of the configurations of the factor scores produced by these two techniques.

This entry presents three coefficients used to evaluate congruence. The first coefficient is called the *coefficient of congruence*. It measures the similarity of two configurations by computing a cosine between matrices of factor scores. The second and third coefficients are the $R_V$ *coefficient* and the *Mantel coefficient*. These two coefficients evaluate the similarity of the *whole configuration* of units. In order to do so, the factor scores of the units are first transformed into a units-by-units square matrix, which reflects the configuration of similarity between the units; and then the similarity between the configurations is measured by a coefficient. For the $R_V$ coefficient, the configuration between the units is obtained by computing a matrix of scalar products between the units, and a cosine between two scalar product matrices evaluates the similarity between two configurations. For the Mantel coefficient, the configuration between the units is obtained by computing a matrix of distance between the units, and a coefficient of correlation between two distance matrices evaluates the similarity between two configurations.

The congruence coefficient was first defined by C. Burt under the name *unadjusted correlation* as

a measure of the similarity of two factorial configurations. The name *congruence coefficient* was later tailored by Ledyard R. Tucker. The congruence coefficient is also sometimes called a *monotonicity coefficient*.

The $R_V$ coefficient was introduced by Yves Escoufier as a measure of similarity between squared symmetric matrices (specifically: positive semidefinite matrices) and as a theoretical tool to analyze multivariate techniques. The $R_V$ coefficient is used in several statistical techniques, such as statis and distatis. In order to compare rectangular matrices with the $R_V$ or the Mantel coefficients, the first step is to transform these rectangular matrices into square matrices.

The Mantel coefficient was originally introduced by Nathan Mantel in epidemiology but it is now widely used in ecology.

The congruence and the Mantel coefficients are cosines (recall that the coefficient of correlation is a *centered* cosine), and as such, they take values between $-1$ and $+1$. The $R_V$ coefficient is also a cosine, but because it is a cosine between two matrices of scalar products (which, technically speaking, are *positive semidefinite* matrices), it corresponds actually to a *squared* cosine, and therefore the $R_V$ coefficient takes values between 0 and 1.

The computational formulas of these three coefficients are almost identical, but their usage and theoretical foundations differ because these coefficients are applied to *different* types of matrices. Also, their sampling distributions differ because of the types of matrices on which they are applied.

## Notations and Computational Formulas

Let $\mathbf{X}$ be an $I$ by $J$ matrix and $\mathbf{Y}$ be an $I$ by $K$ matrix. The **vec** operation transforms a matrix into a vector whose entries are the elements of the matrix. The **trace** operation applies to square matrices and gives the sum of the diagonal elements.

### *Congruence Coefficient*

The congruence coefficient is defined when both matrices have the same number of rows and columns (i.e., $J = K$). These matrices can store factor scores (for observations) or factor loadings (for variables). The congruence coefficient is denoted $\varphi$ or sometimes $r_c$, and it can be computed with three different equivalent formulas (where $^{\mathrm{T}}$ denotes the transpose of a matrix):

$$\varphi = r_c = \frac{\sum_{i,j} x_{i,j} y_{i,j}}{\sqrt{\left(\sum_{i,j} x_{i,j}^2\right)\left(\sum_{i,j} y_{i,j}^2\right)}} \tag{1}$$

$$= \frac{\mathrm{vec}\{\mathbf{X}\}^{\mathrm{T}}\mathrm{vec}\{\mathbf{Y}\}}{\sqrt{\left(\mathrm{vec}\{\mathbf{X}\}^{\mathrm{T}}\mathrm{vec}\{\mathbf{X}\}\right)\left(\mathrm{vec}\{\mathbf{Y}\}^{\mathrm{T}}\mathrm{vec}\{\mathbf{Y}\}\right)}} \tag{2}$$

$$= \frac{\mathrm{trace}\{\mathbf{X}\mathbf{Y}^{\mathrm{T}}\}}{\sqrt{\left(\mathrm{trace}\{\mathbf{X}\mathbf{X}^{\mathrm{T}}\}\right)\left(\mathrm{trace}\{\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\}\right)}} \tag{3}$$

### *$R_V$ Coefficient*

The $R_V$ coefficient was defined by Escoufier as a similarity coefficient between positive semidefinite matrices. Escoufier and Pierre Robert pointed out that the $R_V$ coefficient had important mathematical properties because it can be shown that most multivariate analysis techniques amount to maximizing this coefficient with suitable constraints. Recall, at this point, that a matrix $\mathbf{S}$ is called *positive semidefinite* when it can be obtained as the product of a matrix by its transpose. Formally, we say that $\mathbf{S}$ is positive semidefinite when there exists a matrix $\mathbf{X}$ such that

$$\mathbf{S} = \mathbf{X}\mathbf{X}^{\mathrm{T}}. \tag{4}$$

Note that as a consequence of the definition, positive semidefinite matrices are square and symmetric, and that their diagonal elements are always larger than or equal to zero.

If $\mathbf{S}$ and $\mathbf{T}$ denote two positive semidefinite matrices of same dimensions, the $R_V$ coefficient between them is defined as

$$R_v = \frac{\mathrm{trace}\{\mathbf{S}^{\mathrm{T}}\mathbf{T}\}}{\sqrt{\left(\mathrm{trace}\{\mathbf{S}^{\mathrm{T}}\mathbf{S}\}\right) \times \left(\mathrm{trace}\{\mathbf{T}^{\mathrm{T}}\mathbf{T}\}\right)}} \tag{5}$$

This formula is computationally equivalent to

$$R_V = \frac{\text{vec}\{\mathbf{S}\}^{\mathrm{T}}\text{vec}\{\mathbf{T}\}}{\sqrt{\left(\text{vec}\{\mathbf{S}\}^{\mathrm{T}}\text{vec}\{\mathbf{S}\}\right)\left(\text{vec}\{\mathbf{T}\}^{\mathrm{T}}\text{vec}\{\mathbf{T}\}\right)}} \tag{6}$$

$$= \frac{\sum_{i}^{I}\sum_{j}^{I} s_{i,j}t_{i,j}}{\sqrt{\left(\sum_{i}^{I}\sum_{j}^{I} s_{i,j}^2\right)\left(\sum_{i}^{I}\sum_{j}^{I} t_{i,j}^2\right)}}. \tag{7}$$

For rectangular matrices, the first step is to transform the matrices into positive semidefinite matrices by multiplying each matrix by its transpose. So, in order to compute the value of the $R_V$ coefficient between the $I$ by $J$ matrix $\mathbf{X}$ and the $I$ by $K$ matrix $\mathbf{Y}$, the first step it to compute

$$\mathbf{S} = \mathbf{X}\mathbf{X}^{\mathrm{T}} \text{ and } \mathbf{T} = \mathbf{Y}\mathbf{Y}^{\mathrm{T}}. \tag{8}$$

If we combine Equations 5 and 8, we find that the $R_V$ coefficient between these two rectangular matrices is equal to

$$R_V = \frac{\text{trace}\{\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\}}{\sqrt{\left(\text{trace}\{\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{T}\}\right)\times\left(\text{trace}\{\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\}\right)}}. \tag{9}$$

The comparison of Equations 3 and 9 shows that the congruence and the $R_V$ coefficients are equivalent only in the case of positive semidefinite matrices.

From a linear algebra point of view, the numerator of the $R_V$ coefficient corresponds to a scalar product between positive semidefinite matrices and therefore gives to this set of matrices the structure of a vector space. Within this framework, the denominator of the $R_V$ coefficient is called the *Frobenius* or *Schur* or *Hilbert-Schmidt* matrix scalar product, and the $R_V$ coefficient is a *cosine* between matrices. This vector space structure is responsible for the mathematical properties of the $R_V$ coefficient.

### *Mantel Coefficient*

For the Mantel coefficient, if the data are not already in the form of distances, then the first step is to transform these data into distances. These distances can be Euclidean distances, but any other type of distance will work. If **D** and **B** denote the two $I$ by $I$ distance matrices of interest, then the Mantel coefficient between these two matrices is denoted $r_M$, and it is computed as the coefficient of correlation between their off-diagonal elements as

$$r_M = \frac{\sum_{i=1}^{I-1}\sum_{j=i+1}^{I}(d_{i,j}-\overline{d})(b_{i,j}-\overline{b})}{\sqrt{\left[\sum_{i=1}^{I-1}\sum_{j=i+1}^{I}(d_{i,j}-\overline{d})^2\right]\left[\sum_{i=1}^{I-1}\sum_{j=i+1}^{I}(b_{i,j}-\overline{b})^2\right]}} \tag{10}$$

(where $\overline{d}$ and $\overline{b}$ are the mean of the off-diagonal elements of, respectively, matrices **D** and **B**).

## Tests and Sampling Distributions

The congruence, the $R_V$, and the Mantel coefficients quantify the similarity between two matrices. An obvious practical problem is to be able to perform statistical testing on the value of a given coefficient. In particular it is often important to be able to decide whether a value of coefficient could have been obtained by chance alone. To perform such statistical tests, one needs to derive the sampling distribution of these coefficients under the null hypothesis (i.e., in order to test whether the population coefficient is null). More sophisticated testing requires one to derive the sampling distribution for different values of the population parameters. So far, analytical methods have failed to completely characterize such distributions, but computational approaches have been used with some success. Because the congruence, the $R_V$, and the Mantel coefficients are used with different types of matrices, their sampling distributions differ, and so work done with each type of coefficient has been carried independently of the others.

Some approximations for the sampling distributions have been derived recently for the congruence coefficient and the $R_V$ coefficient, with particular attention given to the $R_V$ coefficient. The sampling distribution for the Mantel coefficient has not been satisfactorily approximated, and the statistical tests provided for this coefficient rely mostly on permutation tests.

### *Congruence Coefficient*

Recognizing that analytical methods were unsuccessful, Bruce Korth and Tucker decided to use Monte Carlo simulations to gain some insights into the sampling distribution of the congruence coefficient. Their work was completed by Wendy J. Broadbooks and Patricia B. Elmore. From this work, it seems that the sampling distribution of the congruence coefficient depends on several parameters, including the original factorial structure and the intensity of the population coefficient, and therefore no simple picture emerges, but some approximations can be used. In particular, for testing that a congruence coefficient is null in the population, an approximate conservative test is to use Fisher's $Z$ transform and to treat the congruence coefficient like a coefficient of correlation. Broadbooks and Elmore have provided tables for population values different from zero. With the availability of fast computers, these tables can easily be extended to accommodate specific cases.

#### Example

Here we use an example from Hervé Abdi and Dominique Valentin (2007). Two wine experts are rating 10 wines on three different scales. The results of their ratings are provided in the two matrices below, denoted **X** and **Y**:

$$\mathbf{X} = \begin{bmatrix} 1 & 6 & 7 \\ 5 & 3 & 2 \\ 6 & 1 & 1 \\ 7 & 1 & 2 \\ 2 & 5 & 4 \\ 3 & 4 & 4 \end{bmatrix} \text{ and } \mathbf{Y} = \begin{bmatrix} 3 & 6 & 7 \\ 4 & 4 & 3 \\ 7 & 1 & 1 \\ 2 & 2 & 2 \\ 2 & 6 & 6 \\ 1 & 7 & 5 \end{bmatrix}. \quad (11)$$

For computing the congruence coefficient, these two matrices are transformed into two vectors of $6 \times 3 = 18$ elements each, and a cosine (cf. Equation 1) is computed between these two vectors. This gives a value of the coefficient of congruence of $\varphi = 7381$. In order to evaluate whether this value is significantly different from zero, a permutation test with 10,000 permutations was performed. In this test, the *rows* of one of the matrices were randomly permuted, and the coefficient of congruence was computed for each of these 10,000 permutations. The probability of obtaining a value of $\varphi = .7381$ under the null hypothesis was evaluated as the proportion of the congruence coefficients larger than $\varphi = .7381$. This gives a value of $p = .0259$, which is small enough to reject the null hypothesis at the .05 alpha level, and thus one can conclude that the agreement between the ratings of these two experts cannot be attributed to chance.

### *$R_V$ Coefficient*

Statistical approaches for the $R_V$ coefficient have focused on permutation tests. In this framework, the permutations are performed on the entries of each column of the rectangular matrices **X** and **Y** used to create the matrices **S** and **T** or directly on the rows *and* columns of **S** and **T**. It is interesting to note that work by Frédérique Kazi-Aoual and colleagues has shown that the mean and the variance of the permutation test distribution can be approximated directly from **S** and **T**.

The first step is to derive an index of the dimensionality or rank of the matrices. This index, denoted $\beta_{\mathbf{S}}$ (for matrix $\mathbf{S} = \mathbf{X}\mathbf{X}^{\mathrm{T}}$), is also known as $\upsilon$ in the brain imaging literature, where it is called a *sphericity* index and is used as an estimation of the number of degrees of freedom for multivariate tests of the general linear model. This index depends on the eigenvalues of the **S** matrix, denoted ${}_{\mathbf{S}}\lambda_\ell$, and is defined as

$$\beta_{\mathbf{S}} = \frac{\left(\sum_{\ell}^{L} {}_{\mathbf{S}}\lambda_\ell\right)^2}{\sum_{\ell}^{L} {}_{\mathbf{S}}\lambda_\ell^2} = \frac{\mathrm{trace}\{\mathbf{S}\}^2}{\mathrm{trace}\{\mathbf{SS}\}}. \quad (12)$$

The mean of the set of permutated coefficients between matrices **S** and **T** is then equal to

$$E(R_V) = \frac{\sqrt{\beta_S \beta_T}}{I-1}. \quad (13)$$

The case of the variance is more complex and involves computing three preliminary quantities for each matrix. The first quantity denoted $\delta_S$ is (for matrix **S**) equal to

$$\delta_S = \frac{\sum_{i}^{I} s_{i,i}^2}{\sum_{\ell}^{L} {}_s\lambda_\ell^2}. \quad (14)$$

The second one is denoted $\alpha_S$ for matrix **S** and is defined as

$$\alpha_S = I - 1 - \beta_S. \quad (15)$$

The third one is denoted $C_S$ (for matrix **S**) and is defined as

$$C_S = \frac{(I-1)[I(I+1)\delta_S - (I-1)(\beta_S + 2)]}{\alpha_S(I-3)}. \quad (16)$$

With these notations, the variance of the permuted coefficients is obtained as

$$V(R_V) = \alpha_S \alpha_T \times \frac{2I(I-1) + (I-3)C_S C_T}{I(I+1)(I-2)(I-1)^3}. \quad (17)$$

For very large matrices, the sampling distribution of the permutated coefficients is relatively similar to a normal distribution (even though it is, in general, *not* normal), and therefore one can use a $Z$ criterion to perform null hypothesis testing or to compute confidence intervals. For example, the criterion

$$Z_{R_V} = \frac{R_V - E(R_V)}{\sqrt{V(R_V)}} \quad (18)$$

can be used to test the null hypothesis that the observed value of $R_V$ was due to chance.

The problem of the lack of normality of the permutation-based sampling distribution of the $R_V$ coefficient has been addressed by Moonseong Heo and K. Ruben Gabriel, who have suggested "normalizing" the sampling distribution by using a log transformation. Recently Julie Josse, Jerome Pagès, and François Husson have refined this approach and indicated that a gamma distribution would give an even better approximation.

### Example

As an example, we use the two scalar product matrices obtained from the matrices used to illustrate the congruence coefficient (cf. Equation 11). For the present example, these original matrices are centered (i.e., the mean of each column has been subtracted from each element of the column) prior to computing the scalar product matrices. Specifically, if $\overline{\mathbf{X}}$ and $\overline{\mathbf{Y}}$ denote the centered matrices derived from **X** and **Y**, we obtain the following scalar product matrices:

$$\mathbf{S} = \overline{\mathbf{X}}\overline{\mathbf{X}}^T = \begin{bmatrix} 29.56 & -8.78 & -20.78 & -20.11 & 12.89 & 7.22 \\ -8.78 & 2.89 & 5.89 & 5.56 & -3.44 & -2.11 \\ -20.78 & 5.89 & 14.89 & 14.56 & -9.44 & -5.11 \\ -20.11 & 5.56 & 14.56 & 16.22 & -10.78 & -5.44 \\ 12.89 & -3.44 & -9.44 & -10.78 & 7.22 & 3.56 \\ 7.22 & -2.11 & -5.11 & -5.44 & 3.56 & 1.89 \end{bmatrix} \quad (19)$$

and

$$\mathbf{T} = \overline{\mathbf{Y}}\overline{\mathbf{Y}}^T = \begin{bmatrix} 11.81 & -3.69 & -15.19 & -9.69 & 8.97 & 7.81 \\ -3.69 & 1.81 & 7.31 & 1.81 & -3.53 & -3.69 \\ -15.19 & 7.31 & 34.81 & 9.31 & -16.03 & -20.19 \\ -9.69 & 1.81 & 9.31 & 10.81 & -6.53 & -5.69 \\ 8.97 & -3.53 & -16.03 & -6.53 & 8.14 & 8.97 \\ 7.81 & -3.69 & -20.19 & -5.69 & 8.97 & 12.81 \end{bmatrix}. \quad (20)$$

We find the following value for the $R_V$ coefficient:

$$
\begin{aligned}
R_V &= \frac{\sum_{i}^{I}\sum_{j}^{I} s_{i,j}t_{i,j}}{\sqrt{\left(\sum_{i}^{I}\sum_{j}^{I} s_{i,j}^2\right)\left(\sum_{i}^{I}\sum_{j}^{I} t_{i,j}^2\right)}} \\
&= \frac{(29.56 \times 11.81) + (-8.78 \times -3.69) + \cdots + (1.89 \times 12.81)}{\sqrt{\left[(29.56) + (-8.78)^2 + \cdots + (1.89)^2\right]\left[(11.81)^2 + (-3.69)^2 + \cdots + (12.81)^2\right]}} \\
&= .7936\ .
\end{aligned}
\tag{21}
$$

To test the significance of a value of $R_V = .7936$, we first compute the following quantities:

$$
\begin{aligned}
\beta_S &= 1.0954 \quad \alpha_S = 3.9046 \\
\delta_S &= 0.2951 \quad C_S - 1.3162 \\
\beta_T &= 1.3851 \quad \alpha_T = 3.6149 \\
\delta_T &= 0.3666 \quad C_T = -0.7045
\end{aligned}
\tag{22}
$$

Plugging these values into Equations 13, 17, and 18, we find

$$
\begin{aligned}
E(R_V) &= 0.2464, \\
V(R_V) &= 0.0422, \text{ and} \\
Z_{R_V} &= 2.66.
\end{aligned}
\tag{23}
$$

Assuming a normal distribution for the $Z_{R_V}$ gives a $p$ value of .0077, which would allow for the rejection of the null hypothesis for the observed value of the $R_V$ coefficient.

#### Permutation Test

As an alternative approach to evaluate whether the value of $R_V = .7936$ is significantly different from zero, a permutation test with 10,000 permutations was performed. In this test, the whole set of rows *and* columns (i.e., the *same* permutation of $I$ elements is used to permute rows and columns) of one of the scalar product matrices was randomly permuted, and the $R_V$ coefficient was computed for each of these 10,000 permutations. The probability of obtaining a value of $R_V = .7936$ under the null hypothesis was evaluated as the proportion of the $R_V$ coefficients larger than $R_V = .7936$. This gave a value of $p = .0281$, which is small enough to reject the null hypothesis at the .05 alpha level. It is worth noting that the normal approximation gives a more liberal (i.e., smaller) value of $p$ than does the nonparametric permutation test (which is more accurate in this case because the sampling distribution of $R_V$ is not normal).

### *Mantel Coefficient*

The exact sampling distribution of the Mantel coefficient is not known. Numerical simulations suggest that, when the distance matrices originate from different independent populations, the sampling distribution of the Mantel coefficient is symmetric (though not normal) with a zero mean. In fact, Mantel, in his original paper, presented some approximations for the variance of the sampling distributions of $r_M$ (derived from the permutation test) and suggested that a normal approximation could be used, but the problem is still open. In practice, though, the probability associated to a specific value of $r_M$ is derived from permutation tests.

#### Example

As an example, two distance matrices derived from the congruence coefficient example (cf. Equation 11) are used. These distance matrices can be computed directly from the scalar product matrices used to illustrate the computation of the $R_V$ coefficient (cf. Equations 19 and 20). Specifically, if **S** is a scalar product matrix and if **s** denotes the vector containing the diagonal elements of **S**, and if 1 denotes an $I$ by 1 vector of ones, then the matrix **D**

of the squared Euclidean distances between the elements of **S** is obtained as (cf. Equation 4):

$$\mathbf{D} = \mathbf{1}\mathbf{s}^{\mathrm{T}} + \mathbf{s}\mathbf{1}^{\mathrm{T}} - \mathbf{2}\mathbf{S}\,. \tag{24}$$

Using Equation 24, we transform the scalar-product matrices from Equations 19 and 20 into the following distance matrices:

$$\mathbf{D} = \begin{bmatrix} 0 & 50 & 86 & 86 & 11 & 17 \\ 50 & 0 & 6 & 8 & 17 & 9 \\ 86 & 6 & 0 & 2 & 41 & 27 \\ 86 & 8 & 2 & 0 & 45 & 29 \\ 11 & 17 & 41 & 45 & 0 & 2 \\ 17 & 9 & 27 & 29 & 2 & 0 \end{bmatrix} \tag{25}$$

and

$$\mathbf{T} = \begin{bmatrix} 0 & 21 & 77 & 42 & 2 & 9 \\ 21 & 0 & 22 & 9 & 17 & 22 \\ 77 & 22 & 0 & 27 & 75 & 88 \\ 42 & 9 & 27 & 0 & 32 & 35 \\ 2 & 17 & 75 & 32 & 0 & 3 \\ 9 & 22 & 88 & 35 & 3 & 0 \end{bmatrix}. \tag{26}$$

For computing the Mantel coefficient, the upper diagonal elements of each of these two matrices are stored into a vector of $\frac{1}{2}I \times (I-1) = 15$ elements, and the standard coefficient of correlation is computed between these two vectors. This gives a value of the Mantel coefficient of $r_M = .5769$. In order to evaluate whether this value is significantly different from zero, a permutation test with 10,000 permutations was performed. In this test, the whole set of rows *and* columns (i.e., the *same* permutation of $I$ elements is used to permute rows and columns) of one of the matrices was randomly permuted, and the Mantel coefficient was computed for each of these 10,000 permutations. The probability of obtaining a value of $r_M = .5769$. under the null hypothesis was evaluated as the proportion of the Mantel coefficients larger than $r_M = .5769$. This gave a value of $p = .0265$, which is small enough to reject the null hypothesis at the .05 alpha level.

## Conclusion

The congruence, $R_V$, and Mantel coefficients all measure slightly different aspects of the notion of congruence. The congruence coefficient is sensitive to the pattern of similarity of the columns of the matrices and therefore will not detect similar configurations when one of the configurations is rotated or dilated. By contrast, both the $R_V$ coefficient and the Mantel coefficients are sensitive to the whole configuration and are insensitive to changes in configuration that involve rotation or dilatation. The $R_V$ coefficient has the additional merit of being theoretically linked to most multivariate methods and of being the base of Procrustes methods such as statis or distatis.

*Hervé Abdi*

***See also*** Coefficients of Correlation, Alienation, and Determination; Principal Components Analysis; $R^2$; Sampling Distributions

### Further Readings

Abdi, H. (2003). Multivariate analysis. In M. Lewis-Beck, A. Bryman, & T. Futing (Eds.), *Encyclopedia for research methods for the social sciences*. Thousand Oaks, CA: Sage.

Abdi, H. (2007). $R_V$ coefficient and congruence coefficient. In N. J. Salkind (Ed.), *Encyclopedia of measurement and statistics*. Thousand Oaks, CA: Sage.

Bedeian, A. G., Armenakis, A. A., & Randolph, W. A. (1988). The significance of congruence coefficients: A comment and statistical test. *Journal of Management, 14,* 559–566.

Borg, I., & Groenen, P. (1997). *Modern multidimensional scaling*. New York: Springer Verlag.

Broadbooks, W. J., & Elmore, P. B. (1987). A Monte Carlo study of the sampling distribution of the congruence coefficient. *Educational & Psychological Measurement, 47,* 1–11.

Burt, C. (1948). Factor analysis and canonical correlations. *British Journal of Psychology, Statistical Section, 1,* 95–106.

Escoufier, Y. (1973). Le traitement des variables vectorielles [Treatment of variable vectors]. *Biometrics, 29,* 751–760.

Harman, H. H. (1976). *Modern factor analysis* (3rd ed. rev.). Chicago: Chicago University Press.

Heo, M., & Gabriel, K. R. (1998). A permutation test of association between configurations by means of the $R_v$ coefficient. *Communications in Statistics Simulation & Computation, 27,* 843–856.

Holmes, S. (2007). Multivariate analysis: The French way. In *Probability and statistics: Essays in honor of David A. Freedman* (Vol. 2, pp. 219–233). Beachwood, OH: Institute of Mathematical Statistics.

Horn, R. A., & Johnson, C. R. (1985). *Matrix analysis.* Cambridge: Cambridge University Press.
Korth, B. A., & Tucker, L. R. (1975). The distribution of chance congruence coefficients from simulated data. *Psychometrika, 40,* 361–372.
Manly, B. J. F. (1997). *Randomization, bootstrap and Monte Carlo methods in biology* (2nd ed.). London: Chapman and Hall.
Schlich, P. (1996). Defining and validating assessor compromises about product distances and attribute correlations. In T. Näs & E. Risvik (Eds.), *Multivariate analysis of data in sensory sciences* (pp. 259–306). New York: Elsevier.
Smouse, P. E., Long, J. C., & Sokal, R. R. (1986). Multiple regression and correlation extensions of the Mantel test of matrix correspondence. *Systematic Zoology, 35,* 627–632.
Worsley, K. J., & Friston, K. J. (1995). Analysis of *f*MRI time-series revisited–Again. *NeuroImage, 2,* 173–181.

# Construct Validity

Construct validity refers to whether the scores of a test or instrument measure the distinct dimension (construct) they are intended to measure. The present entry discusses origins and definitions of construct validation, methods of construct validation, the role of construct validity evidence in the validity argument, and unresolved issues in construct validity.

## Origins and Definitions

Construct validation generally refers to the collection and application of validity evidence intended to support the interpretation and use of test scores as measures of a particular construct. The term *construct* denotes a distinct dimension of individual variation, but use of this term typically carries the connotation that the construct does not allow for direct observation but rather depends on indirect means of measurement. As such, the term *construct* differs from the term *variable* with respect to this connotation. Moreover, the term *construct* is sometimes distinguished from the term *latent variable* because construct connotes a substantive interpretation typically embedded in a body of substantive theory. In contrast, the term *latent variable* refers to a dimension of variability included in a statistical model with or without a clear substantive or theoretical understanding of that dimension and thus can be used in a purely statistical sense. For example, the *latent traits* in item response theory analysis are often introduced as latent variables but not associated with a particular construct until validity evidence supports such an association.

The object of validation has evolved with validity theory. Initially, validation was construed in terms of the validity of a test. Lee Cronbach and others pointed out that validity depends on how a test is scored. For example, detailed content coding of essays might yield highly valid scores whereas general subjective judgments might not. As a result, validity theory shifted its focus from validating tests to validating test scores. In addition, it became clear that the same test scores could be used in more than one way and that the level of validity could vary across uses. For example, the same test scores might offer a highly valid measure of intelligence but only a moderately valid indicator of attention deficit/hyperactivity disorder. As a result, the emphasis of validity theory again shifted from test scores to test score interpretations. Yet a valid interpretation often falls short of justifying a particular use. For example, an employment test might validly measure propensity for job success, but another available test might do as good a job at the same cost but with less adverse impact. In such an instance, the validity of the test score interpretation for the first test would not justify its use for employment testing. Thus, Samuel Messick has urged that test scores are rarely interpreted in a vacuum as a purely academic exercise but are rather collected for some purpose and put to some use. However, in common parlance, one frequently expands the notion of test to refer to the entire procedure of collecting test data (testing), assigning numeric values based on the test data (scoring), making inferences about the level of a construct on the basis of those scores (interpreting), and applying those inferences to practical decisions (use). Thus the term *test validity* lives on as shorthand for the validity of test score interpretations and uses.

Early on, tests were thought to divide into two types: signs and samples. If a test was interpreted as a sign of something else, the something else was understood as a construct, and construct validation was deemed appropriate. For example, responses to items on a personality inventory might be viewed as

signs of personality characteristics, in which case the personality characteristic constitutes the construct of interest. In contrast, some tests were viewed as only samples and construct validation was not deemed necessary. For example, a typing test might sample someone's typing and assess its speed and accuracy. The scores on this one test (produced from a sampling of items that could appear on a test) were assumed to generalize merely on the basis of statistical generalization from a sample to a population. Jane Loevinger and others questioned this distinction by pointing out that the test sample could never be a random sample of all possible exemplars of the behavior in question. For example, a person with high test anxiety might type differently on a typing test from the way the person types at work, and someone else might type more consistently on a brief test than over a full workday. As a result, interpreting the sampled behavior in terms of the full range of generalization always extends beyond mere statistical sampling to broader validity issues. For this reason, all tests are signs as well as samples, and construct validation applies to all tests.

At one time, test validity was neatly divided into three types: content, criterion, and construct, with the idea that one of these three types of validity applied to any one type of test. However, criterion-related validity depends on the construct interpretation of the criterion, and test fairness often turns on construct-irrelevant variance in the predictor scores. Likewise, content validation may offer valuable evidence in support of the interpretation of correct answers but typically will not provide as strong a line of evidence for the interpretation of incorrect answers. For example, someone might know the mathematical concepts but answer a math word problem incorrectly because of insufficient vocabulary or culturally inappropriate examples. Because all tests involve interpretation of the test scores in terms of what they are intended to measure, construct validation applies to all tests. In contemporary thinking, there is a suggestion that all validity should be of one type, construct validity.

This line of development has led to unified (but not unitary) conceptions of validity that elevate construct validity from one kind of validity among others to the whole of validity. Criterion-related evidence provides evidence of construct validity by showing that test scores relate to other variables (i.e., criterion variables) in the predicted ways. Content validity evidence provides evidence of construct validity because it shows that the test properly covers the intended domain of content related to the construct definition. As such, construct validity has grown from humble origins as one relatively esoteric form of validity to the whole of validity, and it has come to encompass other forms of validity evidence.

Messick distinguished two threats to construct validity. *Construct deficiency* applies when a test fails to measure some aspects of the construct that it should measure. For example, a mathematics test that failed to cover some portion of the curriculum for which it was intended would demonstrate this aspect of poor construct validity. In contrast, *construct-irrelevant variance* involves things that the test measures that are not related to the construct of interest and thus should not affect the test scores. The example of a math test that is sensitive to vocabulary level illustrates this aspect. A test with optimal construct validity therefore measures everything that it should measure but nothing that it should not.

Traditionally, validation has been directed toward a specific test, its scores, and their intended interpretation and use. However, construct validation increasingly conceptualizes validation as continuous with extended research programs into the construct measured by the test or tests in question. This shift reflects a broader shift in the behavioral sciences away from *operationalism,* in which a variable is theoretically defined in terms of a single *operational definition,* in favor of *multioperationalism,* in which a variety of different measures triangulate on the same construct. As a field learns to measure a construct in various ways and learns more about how the construct relates to other variables through evidence collected using these measures, the overall understanding of the construct increases. The stronger this overall knowledge base about the construct, the more confidence one can have in interpreting the scores derived from a particular test as measuring this construct. Moreover, the more one knows about the construct, the more specific and varied are the consequences entailed by interpreting test scores as measures of that construct. As a result, one can conceptualize construct validity as broader than test validity because it involves the collection of evidence to validate theories about the underlying construct as measured by

a variety of tests, rather than merely the interpretation of scores from one particular test.

## Construct Validation Methodology

At its inception, when construct validity was considered one kind of validity appropriate to certain kinds of tests, inspection of patterns of correlations offered the primary evidence of construct validity. Lee Cronbach and Paul Meehl described a *nomological net* as a pattern of relationships between variables that partly fixed the meaning of a construct. Later, factor analysis established itself as a primary methodology for providing evidence of construct validity. Loevinger described a *structural aspect* of construct validity as the pattern of relationships between items that compose a test. Factor analysis allows the researcher to investigate the internal structure of item responses, and some combination of replication and confirmatory factor analysis allows the researcher to test theoretical hypotheses about that structure. Such hypotheses typically involve multiple dimensions of variation tapped by items on different subscales and therefore measuring different constructs. A higher order factor may reflect a more general construct that comprises these subscale constructs.

Item response theory typically models dichotomous or polytomous item responses in relation to an underlying latent trait. Although item response theory favors the term *trait,* the models apply to all kinds of constructs. Historically, the emphasis with item response theory has been much more heavily on unidimensional measures and providing evidence that items in a set all measure the same dimension of variation. However, recent developments in factor analysis for dichotomous and polytomous items, coupled with expanded interest in multidimensional item response theory, have brought factor analysis and item response theory together under one umbrella. Item response theory models are generally equivalent to a factor analysis model with a threshold at which item responses change from one discrete response to another based on an underlying continuous dimension. Both factor analysis and item response theory depend on a shared assumption of *local independence,* which means that if one held constant the underlying latent variable, the items would no longer have any statistical association between them. Latent class analysis offers a similar measurement model based on the same basic assumption but applicable to situations in which the latent variable is itself categorical. All three methods typically offer tests of goodness of fit based on the assumption of local independence and the ability of the modeled latent variables to account for the relationships among the item responses.

An important aspect of the above types of evidence involves the separate analysis of various scales or subscales. Analyzing each scale separately does not provide evidence as strong as does analyzing them together. This is because separate analyses work only with local independence of items on the same scale. Analyzing multiple scales combines this evidence with evidence based on relationships between items on different scales. So, for example, three subscales might each fit a one-factor model very well, but a three-factor model might fail miserably when applied to all three sets of items together. Under a hypothetico-deductive framework, testing the stronger hypothesis of multiconstruct local independence offers more support to interpretations of sets of items that pass it than does testing a weaker piecemeal set of hypotheses.

The issue just noted provides some interest in returning to the earlier notion of a nomological net as a pattern of relationships among variables in which the construct of interest is embedded. The idea of a nomological net arose during a period when causation was suspect and laws (i.e., nomic relationships) were conceptualized in terms of patterns of association. In recent years, causation has made a comeback in the behavioral sciences, and methods of modeling networks of causal relations have become more popular. Path analysis can be used to test hypotheses about how a variable fits into such a network of observed variables, and thus path analysis provides construct validity evidence for test scores that fit into such a network as predicted by the construct theory. Structural equation models allow the research to combine both ideas by including both *measurement models* relating items to latent variables (as in factor analysis) and *structural models* that embed the latent variables in a causal network (as in path analysis). These models allow researchers to test complex hypotheses and thus provide even stronger forms of construct validity evidence. When applied to passively observed data, however, such causal models

contain no magic formula for spinning causation out of correlation. Different models will fit the same data, and the same model will fit data generated by different causal mechanisms. Nonetheless, such models allow researchers to construct highly falsifiable hypotheses from theories about the construct that they seek to measure.

Complementary to the above, experimental and quasi-experimental evidence also plays an important role in assessing construct validity. If a test measures a given construct, then efforts to manipulate the value of the construct should result in changes in test scores. For example, consider a standard program evaluation study that demonstrates a causal effect of a particular training program on performance of the targeted skill set. If the measure of performance is well validated and the quality of the training is under question, then this study primarily provides evidence in support of the training program. In contrast, however, if the training program is well validated but the performance measure is under question, then the same study primarily provides evidence in support of the construct validity of the measure. Such evidence can generally be strengthened by showing that the intervention affects the variables that it should but also does not affect the variables that it should not. Showing that a test is responsive to manipulation of a variable that should not affect it offers one way of demonstrating construct-irrelevant variance. For example, admissions tests sometimes provide information about test-taking skills in an effort to minimize the responsiveness of scores to further training in test taking.

Susan Embretson distinguished *construct representation* from *nomothetic span*. The latter refers to the external patterns of relationships with other variables and essentially means the same thing as nomological net. The former refers to the cognitive processes involved in answering test items. To the extent that answering test items involves the intended cognitive processes, the construct is properly represented, and the measurements have higher construct validity. As a result, explicitly modeling the cognitive operation involved in answering specific item types has blossomed as a means of evaluating construct validity, at least in areas in which the underlying cognitive mechanisms are well understood. As an example, if one has a strong construct theory regarding the cognitive processing involved, one can manipulate various cognitive subtasks required to answer items and predict the difficulty of the resulting items from these manipulations.

## Role in Validity Arguments

Modern validity theory generally structures the evaluation of validity on the basis of various strands of evidence in terms of the construction of a *validity argument*. The basic idea is to combine all available evidence into a single argument supporting the intended interpretation and use of the test scores. Recently, Michael Kane has distinguished an *interpretive argument* from the validity argument. The interpretive argument spells out the assumptions and rationale for the intended interpretation of the scores, and the validity argument supports the validity of the interpretive argument, particularly by providing evidence in support of key assumptions. For example, an interpretive argument might indicate that an educational performance mastery test assumes prior exposure and practice with the material. The validity argument might then provide evidence that given these assumptions, test scores correspond to the degree of mastery.

The key to developing an appropriate validity argument rests with identifying the most important and controversial premises that require evidential support. Rival hypotheses often guide this process. The two main threats to construct validity described above yield two main types of rival hypotheses addressed by construct validity evidence. For example, sensitivity to transient emotional states might offer a rival hypothesis to the validity of a personality scale related to construct-irrelevant variance. *Differential item functioning*, in which test items relate to the construct differently for different groups of test takers, also relates to construct-irrelevant variance, yielding rival hypotheses about test scores related to group characteristics. A rival hypothesis that a clinical depression inventory captures only one aspect of depressive symptoms involves a rival hypothesis about construct deficiency.

## Unresolved Issues

A central controversy in contemporary validity theory involves the disagreements over the breadth

of validity evidence. Construct validation provides an integrative framework that ties together all forms of validity evidence in a way continuous with empirical research into the construct, but some have suggested a less expansive view of validity as more practical. Construct validity evidence based on test consequences remains a continuing point of controversy, particularly with respect to the notion of *consequential validity* as a distinct form of validity. Finally, there remains a fundamental tension in modern validity theory between the traditional fact–value dichotomy and the fundamental role of values and evaluation in assessing the evidence in favor of specific tests, scores, interpretations, and uses.

*Keith A. Markus and Chia-ying Lin*

***See also*** Content Validity; Criterion Validity; Structural Equation Modeling

**Further Readings**

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education and Praeger.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education and Macmillan.

Wainer, H., & Braun, H. I. (1988). *Test validity.* Mahwah, NJ: Lawrence Erlbaum.

# Content Analysis

Content analysis is a research technique for making replicable and valid inferences from texts (or other meaningful matter) to the context of their use.

This entry further explores the definition and conceptions of content analysis. It then provides information on its conceptual framework and the steps involved in performing content analysis.

## Definition and Conception

The phrase *content analysis,* first mentioned in a 1941 paper by Douglas Waples and Bernard Berelson, became defined in 1948 by Paul F. Lazarsfeld and Berelson. *Webster's Dictionary* has listed *content analysis* since its 1961 edition. However, the practice of analyzing media matter is almost as old as writing. It became of interest to the church, worried about the effects of the written word other than God's; to governments, trying to settle political, legal, and religious disputes; to journalists, hoping to document the changes in newspaper publishing due to its commercialization and popularization; to corporations interested in surveying their symbolic environments for opportunities and threats; and to social scientists, originally drawn into the competition between the press and newly emerging media, then radio and television, but soon discovering the importance of all kinds of mediated communication to understand social, political, economic, and psychological phenomena. Communication research advanced content analysis, but owing to the proliferation of media and the recognition that humans define themselves and each other in communication, coordinate their beliefs and actions in communication, and construct the realities they live with in communication, content analysis is now used by literally all social sciences.

As a *technique,* content analysis embraces specialized procedures. It is teachable. Its use can be divorced from the authority of the researcher. As a *research* technique, content analysis can provide new kinds of understanding social phenomena or inform decisions on pertinent actions. Content analysis is a scientific tool.

All techniques are expected to be reliable. Scientific research techniques should result in replicable findings. Replicability requires research procedures to be explicit and communicable so that researchers, working at different times and perhaps under different circumstances, can apply them and come to the same conclusions about the same phenomena.

Scientific research must also yield valid results. To establish validity, research results must survive in the face of independently available evidence of what they claim. The methodological requirements of reliability and validity are not unique to content

analysis but make particular demands on the technique that are not found as problematic in other methods of inquiry.

The reference to *text* is not intended to restrict content analysis to written material. The parenthetical phrase "or other meaningful matter" is to imply content analysis's applicability to anything humanly significant: images, works of art, maps, signs, symbols, postage stamps, songs, and music, whether mass produced, created in conversations, or private. Texts, whether composed by individual authors or produced by social institutions, are always intended to point their users to something beyond their physicality. However, content analysis does not presume that readers read a text as intended by its source; in fact, authors may be quite irrelevant, often unknown. In content analysis, available texts are analyzed to answer research questions not necessarily shared by everyone.

What distinguishes content analysis from most observational methods in the social sciences is that *the answers to its research questions are inferred from available text*. Content analysts are not interested in the physicality of texts that can be observed, measured, and objectively described. The alphabetical characters of written matter, the pixels of digital images, and the sounds one can manipulate at a control panel are mere vehicles of communication. What text means to somebody, what it represents, highlights and excludes, encourages or deters—all these phenomena do not reside inside a text but come to light in processes of someone's reading, interpreting, analyzing, concluding, and in the case of content analysis, answering pertinent research questions concerning the text's context of use.

Typical research questions that content analysts might answer are, What are the consequences for heavy and light viewers of exposure to violent television shows? What are the attitudes of a writer on issues not mentioned? Who is the author of an anonymously written work? Is a suicide note real, requiring intervention, or an empty threat? Which of two textbooks is more readable by sixth graders? What is the likely diagnosis for a psychiatric patient, known through an interview or the responses to a Rorschach test? What is the ethnic, gender, or ideological bias of a newspaper? Which economic theory underlies the reporting of business news in the national press? What is the likelihood of cross-border hostilities as a function of how one country's national press portrays its neighbor? What are a city's problems as inferred from citizens' letters to its mayor? What do school children learn about their nation's history through textbooks? What criteria do Internet users employ to authenticate electronic documents?

## Other Conceptions

Unlike content analysis, observation and measurement go directly to the phenomenon of analytic interest. Temperature and population statistics describe tangible phenomena. Experiments with human participants tend to define the range of responses in directly analyzable form, just as structured interviews delineate the interviewees' multiple choices among answers to prepared interview questions. Structured interviews and experiments with participants acknowledge subjects' responses to meanings but bypass them by standardization. Content analysts struggle with unstructured meanings.

Social scientific literature does contain conceptions of content analysis that mimic observational methods, such as those of George A. Miller, who characterizes content analysis as a method for putting large numbers of units of verbal matter into analyzable categories. A definition of this kind provides no place for methodological standards. Berelson's widely cited definition fares not much better. For him, "content analysis is a research technique for the objective, systematic and quantitative description of the manifest content of communication" (p. 18). The restriction to manifest content would rule out content analyses of psychotherapeutic matter or of diplomatic exchanges, both of which tend to rely on subtle clues to needed inferences. The requirement of quantification, associated with objectivity, has been challenged, especially because the reading of text is qualitative to start and interpretive research favors qualitative procedures without being unscientific. Taking the questionable attributes out of Berelson's definition reduces content analysis to the systematic analysis of content, which relies on a metaphor of content that locates the object of analysis inside the text—a conception that some researchers believe is not only misleading but also prevents the formulation of sound methodology.

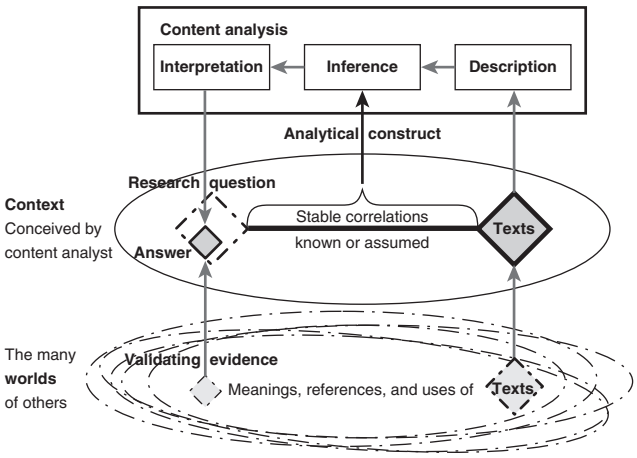



**Figure 1** A Framework for Content Analysis

There are definitions, such as Charles Osgood's or Ole Holsti's, that admit inferences but restrict them to the source or destination of the analyzed messages. These definitions provide for the use of validity criteria by allowing independent evidence to be brought to bear on content analysis results, but they limit the analysis to causal inferences.

## Conceptual Framework

Figure 1 depicts the methodologically relevant elements of content analysis. A content analysis usually starts with either or both (a) available text that, on careful reading, poses scientific research questions or (b) research questions that lead the researcher to search for texts that could answer them.

### *Research Questions*

In content analysis, research questions need to go outside the physicality of text into the world of others. The main motivation for using content analysis is that the answers sought cannot be found by direct observation, be it because the phenomena of interest are historical, hence past; enshrined in the mind of important people, not available for interviews; concern policies that are deliberately hidden, as by wartime enemies; or concern the anticipated effects of available communications, hence not yet present. If these phenomena were observable directly, content analysis of texts would be redundant. Content analysts pursue questions that could conceivably be answered by examining texts but that could also be validated by other means, at least in principle.

The latter rules out questions that have to do with a researcher's skill in processing text, albeit systematically. For example, the question of how much violence is featured on television is answerable by counting incidences of it. For content analysts, an index of violence on television needs to have empirical validity in the sense that it needs to say something about how audiences of violent shows react, how their conception of the world is shaped by being exposed to television violence, or whether it encourages or discourages engaging in violent acts. Inferences about antecedents and consequences can be validated, at least in principle. Counts can be validated only by recounting. Without designating where validating evidence could be found, statements about the physicality of text would not answer the research questions that define a content analysis.

Research questions must admit alternative answers. They are similar to a set of hypotheses to be tested, except that inferences from text determine choices among them.

### *Context of the Analysis*

All texts can be read in multiple ways and provide diverging information to readers with diverging competencies and interests. Content analysts are not different in this regard, except for their mastery of analytical techniques. To keep the range of possible inferences manageable, content analysts need to construct a context in which their research questions can be related to available texts in ways that are transparent and available for examination by fellow scientists. This restriction is quite natural. Psychologists construct their world unlike sociologists do, and what is relevant when policy recommendation or evidence in court needs to be provided may have little to do with an analysis that aims at deciding when different parts of the Bible were written. The context of a content analysis always is the analyst's choice. There are no restrictions except for having to be explicit and arguably related to the world of others for whom the analyzed text means something, refers to

something, and is useful or effective, though not necessarily as content analysts conceptualize these things.

## Description of Text

Usually, the first step in a content analysis is a description of the text. Mary Bock called content analyses that stop there "impressionistic" because they leave open what a description could mean. Three types of description may be distinguished: (1) selected word counts, (2) categorizations by common dictionaries or thesauri, and (3) recording or scaling by human coders.

### *Selected Word Counts*

Selected word counts can easily be obtained mechanically and afford numerous comparisons by sources or situations or over time. For example, the 12 most frequent words uttered by Paris Hilton in an interview with Larry King were 285 *I,* 66 *you,* 61 *my,* 48 *like,* 45 *yes,* 44 *really,* 40 *me,* 33 *I'm,* 32 *people,* 28 *they,* 17 *life* and *time,* and 16 *jail.* That *I* is by far the most frequent word may suggest that the interviewee talked largely about herself and her own life, which incidentally included a brief visit in jail. Such a distribution of words is interesting not only because normally one does not think about words when listening to conversations but also because its skewedness is quite unusual and invites explanations. But whether Hilton is self-centered, whether her response was due to Larry King's questioning, how this interview differed from others he conducted, and what the interview actually revealed to the television audience remain speculation. Nevertheless, frequencies offer an alternative to merely listening or observing.

Many computer aids to content analysis start with words, usually omitting function words, such as articles, stemming them by removing grammatical endings, or focusing on words of particular interest. In that process, the textual environments of words are abandoned or, in the case of *keywords in context* lists, significantly reduced.

### *Categorizing by Common Dictionaries or Thesauri*

Categorization by common dictionaries or thesauri is based on the assumptions that (a) textual meanings reside in words, not in syntax and organization; (b) meanings are shared by everyone—"manifest," in Berelson's definition—as implied in the use of published dictionaries and thesauri; and (c) certain differentiations among word meanings can be omitted in favor of the gist of semantic word classes. Tagging texts is standard in several computer aids for content analysis. The General Inquirer software, for example, assigns the words *I, me, mine,* and *myself* to the tag "self" and the tags "self," "selves," and "others" to the second-order tag "person." Where words are ambiguous, such as *play,* the General Inquirer looks for disambiguating words in the ambiguous word's environment—looking, in the case of *play,* for example, for words relating to children and toys, musical instruments, theatrical performances, or work—and thereby achieves a less ambiguous tagging.

Tagging is also used to scale favorable or unfavorable attributes or assign positive and negative signs to references.

### *Recording or Scaling by Human Coders*

Recording or scaling by human coders is the traditional and by far the most common path taken to obtain analyzable descriptions of text. The demand for content analysis to be reliable is met by standard coding instructions, which all coders are asked to apply uniformly to all units of analysis. Units may be words, propositions, paragraphs, news items, or whole publications of printed matter; scenes, actors, episodes, or whole movies in the visual domain; or utterances, turns taken, themes discussed, or decisions made in conversations.

The use of standard coding instructions offers content analysts not only the possibility of analyzing larger volumes of text and employing many coders but also a choice between emic and etic descriptions—*emic* by relying on the very categories that a designated group of readers would use to describe the textual matter, *etic* by deriving coding categories from the theories of the context that the content analysts have adopted. The latter choice enables content analysts to describe latent contents and approach phenomena that ordinary writers and readers may not be aware of. "Good" and "bad" are categories nearly everyone understands alike, but "prosocial" and "antisocial"

attitudes, the concept of framing, or the idea of a numerical strength of word associations needs to be carefully defined, exemplified, and tested for reliability.

## Inference

### *Abduction*

Although sampling considerations are important in selecting texts for analysis, the type of inference that distinguishes content analysis from observational methods is abduction—not induction or deduction. Abduction proceeds from particulars—texts—to essentially different particulars—the answers to research questions. For example, inferring the identity of the author from textual qualities of an unsigned work; inferring levels of anxiety from speech disturbances; inferring a source's conceptualization from the proximity of words it uses; inferring Stalin's successor from public speeches by Politburo members at the occasion of Stalin's birthday; or inferring possible solutions to a conflict entailed by the metaphors used in characterizing that conflict.

### *Analytical Constructs*

Inferences of this kind require some evidential support that should stem from the known, assumed, theorized, or experimentally confirmed stable correlations between the textuality as described and the set of answers to the research question under investigation. Usually, this evidential support needs to be operationalized into a form applicable to the descriptions of available texts and interpretable as answers to the research questions. Such operationalizations can take numerous forms. By intuition, one may equate a measure of the space devoted to a topic with the importance a source attributes to it. The relation between different speech disturbances and the diagnosis of certain psychopathologies may be established by correlation. The relation between the proximity of words and associations, having been experimentally confirmed, may be operation-alized in clustering algorithms that compute word clusters from strings of words.

While the evidential support for the intended inferences can come from anywhere, content analysts cannot bypass justifying this step. It would be methodologically inadmissible to claim to have analyzed "the" content of a certain news channel, as if no inference were made or as if content were contained in its transmissions, alike for everyone, including content analysts. It is equally inadmissible to conclude from applying a standard coding instrument and a sound statistics on reliably coded data, that the results of a content analysis say anything about the many worlds of others. They may represent nothing other than the content analyst's systematized conceptions.

Regarding the analytical construct, content analysts face two tasks, preparatory and applied. Before designing a content analysis, researchers may need to test or explore available evidence, including theories of the stable relations on grounds of which the use of analytical constructs can be justified. After processing the textual data, the inferences tendered will require similar justifications.

## Interpretation

The result of an inference needs to be interpreted so as to select among the possible answers to the given research question. In identifying the author of an unsigned document, one may have to translate similarities between signed and unsigned documents into probabilities associated with conceivable authors. In predicting the use of a weapon system from enemy domestic propaganda, one may have to extrapolate the fluctuations of mentioning it into a set of dates. In ascertaining gender biases in educational material, one may have to transform the frequencies of gender references and their evaluation into weights of one gender over another.

Interpreting inferences in order to select among alternative answers to a research question can be quite rigorous. Merely testing hypotheses on the descriptive accounts of available texts stays within the impressionistic nature of these descriptions and has little to do with content analysis.

## Criteria for Judging Results

There are essentially three conditions for judging the acceptability of content analysis results. In the absence of direct validating evidence for the

inferences that content analysts make, there remain reliability and plausibility.

### *Reliability*

Reliability is the ability of the research process to be replicated elsewhere. It assures content analysts that their data are rooted in shared ground and other researchers that they can figure out what the reported findings mean or add their own data to them. Traditionally, the most unreliable part of a content analysis is the recording, categorization, or scaling of text by human coders, and content analysts employing coders for this purpose are required to assess the reliability of that process quantitatively. Measures of reliability are provided by agreement coefficients with suitable reliability interpretations, such as Scott's $\pi$ (pi) and Krippendorff's $\alpha$ (alpha). The literature contains recommendations regarding the minimum agreement required for an analytical process to be sufficiently reliable. However, that minimum should be derived from the consequences of answering the research question incorrectly. Some disagreements among coders may not make a difference, but others could direct the process to a different result.

### *Plausibility*

Computer content analysts pride themselves in having bypassed reliability problems. However, all content analysts need to establish the plausibility of the path taken from texts to their results. This presupposes explicitness as to the analytical steps taken. The inability to examine critically the steps by which a content analysis proceeded to its conclusion introduces doubts in whether the analysis can be trusted, and implausibility can fail the effort. Content analysts cannot hide behind obscure algorithms whose inferences are unclear. Plausibility may not be quantifiable, as reliability is, but it is one criterion all content analyses must satisfy.

### *Validity*

In content analysis, validity may be demonstrated variously. The preferred validity is *predictive,* matching the answers to the research question with subsequently obtained facts. When direct and post facto validation is not possible, content analysts may need to rely on indirect evidence. For example, when inferring the psychopathology of a historical figure, accounts by the person's contemporaries, actions on record, or comparisons with today's norms may be used to triangulate the inferences. Similarly, when military intentions are inferred from the domestic broadcasts of wartime enemies, such intentions may be correlated with observable consequences or remain on record, allowing validation at a later time. *Correlative* validity is demonstrated when the results of a content analysis correlate with other variables. *Structural* validity refers to the degree to which the analytical construct employed does adequately model the stable relations underlying the inferences, and *functional* validity refers to the history of the analytical construct's successes. *Semantic* validity concerns the validity of the description of textual matter relative to a designated group of readers, and *sampling* validity concerns the representativeness of the sampled text. Unlike in observational research, texts need to be sampled in view of their ability to provide the answers to research questions, not necessarily to represent the typical content produced by their authors.

*Klaus Krippendorff*

***See also*** Hypothesis; Interrater Reliability; Krippendorff's Alpha; Reliability; Validity of Research Conclusions

**Further Readings**

Berelson, B. (1952). *Content analysis in communications research*. New York: Free Press.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2nd ed.). Thousand Oaks, CA: Sage.

Krippendorff, K., & Bock, M. A. (2008). *The content analysis reader.* Thousand Oaks, CA: Sage.

## Content Validity

Content validity refers to the extent to which the items on a test are fairly representative of the entire domain the test seeks to measure. This entry discusses origins and definitions of content validation, methods of content validation, the role of

content validity evidence in validity arguments, and unresolved issues in content validation.

## Origins and Definitions

One of the strengths of content validation is the simple and intuitive nature of its basic idea, which holds that what a test seeks to measure constitutes a *content domain* and the items on the test should sample from that domain in a way that makes the test items representative of the entire domain. Content validation methods seek to assess this quality of the items on a test. Nonetheless, the underlying theory of content validation is fraught with controversies and conceptual challenges.

At one time, different forms of validation, and indeed validity, were thought to apply to different types of tests. Florence Goodenough made an influential distinction between tests that serve as *samples* and tests that serve as *signs*. From this view, personality tests offer the canonical example of tests as signs because personality tests do not sample from a domain of behavior that constitutes the personality variable but rather serve to indicate an underlying personality trait. In contrast, educational achievement tests offer the canonical example of tests as samples because the items sample from a knowledge or skill domain, operationally defined in terms of behaviors that demonstrate that corresponding knowledge or skill that the test measures achievement in. For example, if an addition test contains items representative of all combinations of single digits, then it may adequately represent addition of single-digit numbers, but it would not adequately represent addition of numbers with more than one digit.

Jane Loevinger and others have argued that the above distinction does not hold up because all tests actually function as signs. The inferences drawn from test scores always extend beyond the test-taking behaviors themselves, but it is impossible for the test to include anything beyond test-taking behaviors. Even work samples can extend only to samples of work gathered within the testing procedure (as opposed to portfolios, which lack the standardization of testing procedures). To return to the above example, one does not use an addition test to draw conclusions only about answering addition items on a test but seeks to generalize to the ability to add in contexts outside addition tests.

At the heart of the above issue lies the paradigmatic shift from discrete forms of validity, each appropriate to one kind of test, to a more unified approach to test validation. The term *content validity* initially differentiated one form of validity from *criterion validity* (divisible into concurrent validity and predictive validity, depending on the timing of the collection of the criterion data) and *construct validity* (which initially referred primarily to the pattern of correlations with other variables, the *nomological net,* and to the pattern of association between the scores on individual items within the test). Each type of validity arose from a set of practices that the field developed to address a particular type of practical application of test use. Content validity was the means of validating tests used to sample a content domain and evaluate mastery within that domain. The *unified view of validity* initiated by Jane Loevinger and Lee Cronbach, and elaborated by Samuel Messick, sought to forge a single theory of test validation that subsumed these disparate practices.

The basic practical concern involved the fact that assessment of the representativeness of the content domain achieved by a set of items does not provide a sufficient basis to evaluate the soundness of inferences from scores on the test. For example, a student correctly answering arithmetic items at a level above chance offers stronger support for the conclusion that he or she can do the arithmetic involved than the same student failing to correctly answer the items offers for the conclusion that he or she cannot do the arithmetic. It may be that the student can correctly calculate 6 divided by 2 but has not been exposed to the 6/2 notation used in the test items. In another context, a conscientious employee might be rated low on a performance scale because the items involve tasks that are important to and representative of the domain of conscientious work behaviors, but opportunities for which come up extremely rarely in the course of routine work (e.g., reports defective equipment when encountered). Similarly, a test with highly representative items might have inadequate reliability or other deficiencies that reduce the validity of inferences from its scores. The traditional approach to dividing up types of validity and categorizing tests with respect to the

appropriate type of validation tends in practice to encourage reliance on just one kind of validity evidence for a given test. Because just one type alone, including content-related validation evidence, does not suffice to underwrite the use of a test, the unified view sought to discourage such categorical typologies of either tests or validity types and replace these with validation methods that combined different forms of evidence for the validity of the same test.

As Stephen Sireci and others have argued, the problem with the unified approach with respect to content validation stems directly from this effort to improve on inadequate test validation practices. A central ethos of unified approaches involves the rejection of a simple checklist approach to validation in which completion of a fixed set of steps results in a permanently validated test that requires no further research or evaluation. As an antidote to this checklist conception, Michael Kane and others elaborated the concept of a *validity argument*. The basic idea was that test validation involves building an argument that combines multiple lines of evidence of the overall evaluation of a use or interpretation of scores derived from a test. To avoid a checklist, the argument approach leaves it open to the test validator to exercise judgment and select the lines of evidence that are most appropriate in a given instance. This generally involves selecting the premises of the validation argument that bear the most controversy and for which empirical support can be gathered within practical constraints on what amounts to a reasonable effort. One would not waste resources gathering empirical evidence for claims that no one would question. Similarly, one would not violate ethical standards in order to validate a test of neural functioning by damaging various portions of the cortex in order to experimentally manipulate the variable with random assignment. Nor would one waste resources on an enormous and costly effort to test one assumption if those resources could be better used to test several others in a less costly fashion. In short, the validity argument approach to test validation does not specify that any particular line of evidence is required of a validity argument. As a result, an effort to discourage reliance on content validation evidence alone may have swung the pendulum too far in the opposite direction by opening the door to validation efforts that exclude content validation where it could provide an important and perhaps necessary line of support. These considerations have led to proposals to modify the argument approach to validation in ways that make content-related evidence necessary or at least strongly recommended for tests based on sampling from a content domain.

Contemporary approaches to content validation typically distinguish various aspects of content validity. A clear *domain definition* is foundational for all the other aspects of content validity because without a clear definition of the domain, test developers, test users, or anyone attempting to do validation research has no basis for a clear assessment of the remaining aspects. This aspect of content validation closely relates to the emphasis in the *Standards for Educational and Psychological Testing* on clearly defining the purpose of a test as the first step in test validation.

A second aspect of content validity, *domain relevance,* draws a further connection between content validation and the intended purpose of the test. Once the domain has been defined, domain relevance describes the degree to which the defined domain bears importance to the purpose of the test. For example, one could imagine a test that does a very good job of sampling the skills required to greet visitors, identify whom they wish to see, schedule appointments, and otherwise exercise the judgment and complete the tasks required of an effective receptionist. However, if the test use involves selecting applicants for a back office secretarial position that does not involve serving as a receptionist, then the test would not have good domain relevance for the intended purpose. This aspect of content validation relates to a quality of the defined domain independent of how well the test taps that domain.

In contrast, *domain representation* does not evaluate the defined domain but rather evaluates the effectiveness with which the test samples that domain. Clearly, this aspect of content validation depends on the previous two. Strong content representation does not advance the quality of a test if the items represent a domain with low relevance. Furthermore, even if the items do represent a domain well, the test developer has no effective means of ascertaining that fact without a clear domain definition. Domain representation can

suffer in two ways: Items on the test may fail to sample some portion of the test domain, in which case the validity of the test suffers as a result of construct underrepresentation. Alternatively, the test might contain items from outside the test domain, in which case these items introduce construct-irrelevant variance into the test total score. It is also possible that the test samples all and only the test domain but does so in a way that overemphasizes some areas of the domain while underemphasizing other areas. In such a case, the items sample the entire domain but in a nonrepresentative manner. An example would be an addition test where 75% of the items involved adding only even numbers and no odd numbers.

An additional aspect of content validation involves clear, detailed, and thorough documentation of the test construction procedures. This aspect of content validation reflects the epistemic aspect of modern test validity theory: Even if a test provides an excellent measure of its intended construct, test users cannot justify the use of the test unless they know that the test provides an excellent measure. Test validation involves justifying an interpretation or use of a test, and content validation involves justifying the test domain and the effectiveness with which the test samples that domain. Documentation of the process leading to the domain definition and generation of the item pool provides a valuable source of content-related validity evidence. One primary element of such documentation, the *test blueprint,* specifies the various areas of the test domain and the number of items from each of those areas. Documentation of the process used to construct the test in keeping with the specified test blueprint thereby plays a central role in evaluating the congruency between the test domain and the items on the test.

The earlier passages of this entry have left open the question of whether content validation refers only to the items on the test or also to the processes involved in answering those items. Construct validation has its origins in a time when tests as the object of validation were not yet clearly distinguished from test scores or test score interpretations. As such, most early accounts focused on the items rather than the processes involved in answering them. Understood this way, content validation focuses on qualities of the test rather than qualities of test scores or interpretations. However, as noted above, even this test-centered approach to content validity remains relative to the purpose for which one uses the test. Domain relevance depends on this purpose, and the purpose of the test should ideally shape the conceptualization of the test domain. However, focus on just the content of the items allows for a broadening of content validation beyond the conception of a test as measuring a construct conceptualized as a latent variable representing a single dimension of variation. It allows, for instance, for a test domain that spans a set of tasks linked another way but heterogeneous in the cognitive processes involved in completing them. An example might be the domain of tasks associated with troubleshooting a complex piece of technology such as a computer network. No one algorithm or process might serve to troubleshoot every problem in the domain, but content validation held separate from response processes can nonetheless apply to such a test.

In contrast, the idea that content validity applies to response processes existed as a minority position for most of the history of content validation, but has close affinities to both the unified notion of validation as an overall evaluation based on the sum of the available evidence and also with cognitive approaches to test development and validation. Whereas representativeness of the item content bears more on a quality of the stimulus materials, representativeness of the response processes bears more on an underlying individual differences variable as a property of the person tested. Susan Embretson has distinguished *construct representation,* involving the extent to which items require the cognitive processes that the test is supposed to measure, from *nomothetic span,* which is the extent to which the test bears the expected patterns of association with other variables (what Cronbach and Paul Meehl called *nomological network*). The former involves content validation applied to processes whereas the latter involves methods more closely associated with criterion-related validation and construct validation methods.

## Content Validation Methodology

Content-related validity evidence draws heavily from the test development process. The content domain should be clearly defined at the start of

this process, *item specifications* should be justified in terms of this domain definition, *item construction* should be guided and justified by the item specifications, and the overall test blueprint that assembles the test from the item pool should also be grounded in and justified by the domain definition. Careful documentation of each of these processes provides a key source of validity evidence.

A standard method for assessing content validity involves judgments by *subject matter experts* (SMEs) with expertise in the content of the test. Two or more SMEs rate each item, although large or diverse tests may require different SMEs for different items. Ratings typically involve domain relevance or importance of the content in individual test items. Good items have high means and low standard deviations, indicating high agreement among raters. John Flanagan introduced a *critical incident technique* for generating and evaluating performance-based items. C. H. Lawshe, Lewis Aiken, and Ronald Hambleton each introduced quantitative measures of agreement for use with criterion-related validation research. Victor Martuza introduced a *content validity index,* which has generated a body of research in the nursing literature. A number of authors have also explored multivariate methods for investigating and summarizing SME ratings, including factor analysis and multidimensional scaling methods. Perhaps not surprisingly, the results can be sensitive to the approach taken to structuring the judgment task.

Statistical analysis of item scores can also be used to evaluate content validity by showing that the content domain theory is consistent with the clustering of items into related sets of items by some statistical criteria. These methods include factor analysis, multidimensional scaling methods, and cluster analysis. Applied to content validation, these methods overlap to some degree with construct validation methods directed toward the internal structure of a test. Test developers most often combine such methods with methods based on SME ratings to lessen interpretational ambiguity of the statistical results.

A growing area of test validation related to content involves cognitive approaches to modeling the processes involved in answering specific item types. Work by Embretson and Robert Mislevy exemplifies this approach, and such approaches focus on the construct representation aspect of test validity described above. This methodology relies on a strong cognitive theory of how test takers process test items and thus applies best when item response strategies are relatively well understood and homogeneous across items. The approach sometimes bears a strong relation to the facet analysis methods of Louis Guttman in that item specifications describe and quantify a variety of item attributes, and these can be used to predict features of item response patterns such as item difficulty. This approach bears directly on content validity because it requires a detailed theory relating how items are answered to what the items measure. Response process information can also be useful in extrapolating from the measured content domain to broader inferences in applied testing, as described in the next section.

## Role in Validity Arguments

At one time, the dominant approach was to identify certain tests as the type of test to which content validation applies and rely on content validity evidence for the evaluation of such tests. Currently, few if any scholars would advocate sole reliance on content validity evidence for any test. Instead, content-related evidence joins with other evidence to support key inferences and assumptions in a validity argument that combines various sources of evidence to support an overall assessment of the test score interpretation and use.

Kane has suggested a two-step approach in which one first constructs an argument for test score interpretations and then evaluates that argument with a test validity argument. Kane has suggested a general structure involving four key inferences to which content validity evidence can contribute support. First, the prescribed scoring method involves an inference from observed test-taking behaviors to a specific quantification intended to contribute to measurement through an overall quantitative summary of the test takers' responses. Second, test score interpretation involves generalization from the observed test score to the defined content domain sampled by the test items. Third, applied testing often involves a further inference that extrapolates from the measured content domain to a broader domain of inference that the test does not fully sample. Finally, most applied testing involves a final set of

inferences from the extrapolated level of performance to implications for actions and decisions applied to a particular test taker who earns a particular test score.

Interpretation of statistical models used to provide criterion- and construct-related validity evidence would generally remain indeterminate were it not for the grounding of test score interpretations provided by content-related evidence. While not a fixed foundation for inference, content-related evidence provides a strong basis for taking one interpretation of a nomothetic structure as more plausible than various rival hypotheses. As such, content-related validity evidence continues to play an important role in test development and complements other forms of validity evidence in validity arguments.

### Unresolved Issues

As validity theory continues to evolve, a number of issues in content validation remain unresolved. For instance, the relative merits of restricting content validation to test content or expanding it to involve item response processes warrant further attention. A variety of aspects of content validity have been identified, suggesting a multidimensional attribute of tests, but quantitative assessments of content validity generally emphasize single-number summaries. Finally, the ability to evaluate content validity in real time with computer-adaptive testing remains an active area of research.

*Keith A. Markus and Kellie M. Smith*

***See also*** Construct Validity; Criterion Validity

#### Further Readings

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Crocker, L. M., Miller, D., & Franks, E. A. (1989). Quantitative methods for assessing the fit between test and curriculum. *Applied Measurement in Education, 2,* 179–194.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93,* 179–197.

Kane, M. (2006). Content-related validity evidence in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 131–153). Mahwah, NJ: Lawrence Erlbaum.

McKenzie, J. F., Wood, M. L., Kotecki, J. E., Clark, J. K., & Brey, R. A. (1999). Research notes establishing content validity: Using qualitative and quantitative steps. *American Journal of Health Behavior, 23,* 311–318.

Popham, W. J. (1992). Appropriate expectations for content judgments regarding teacher licensure tests. *Applied Measurement in Education, 5,* 285–301.

Sireci, S. (1998). The construct of content validity. *Social Indicators Research, 45,* 83–117.

## Contrast Analysis

A standard analysis of variance (ANOVA) provides an *F* test, which is called an *omnibus test* because it reflects all possible differences between the means of the groups analyzed by the ANOVA. However, most experimenters want to draw conclusions more precise than "the experimental manipulation has an effect on participants' behavior." Precise conclusions can be obtained from *contrast analysis* because a contrast expresses a specific question about the pattern of results of an ANOVA. Specifically, a contrast corresponds to a prediction precise enough to be translated into a set of numbers called *contrast coefficients,* which reflect the prediction. The correlation between the contrast coefficients and the observed group means directly evaluates the similarity between the prediction and the results.

When performing a contrast analysis, one needs to distinguish whether the contrasts are *planned* or post hoc. Planned, or a priori, contrasts are selected before running the experiment. In general, they reflect the hypotheses the experimenter wants to test, and there are usually *few* of them. Post hoc, or a posteriori (after the fact), contrasts are decided after the experiment has been run. The goal of a posteriori contrasts is to ensure that unexpected results are reliable.

When performing a planned analysis involving several contrasts, one needs to evaluate whether these contrasts are mutually orthogonal or not. Two contrasts are *orthogonal* when their contrast coefficients are uncorrelated (i.e., their coefficient

of correlation is zero). The number of possible orthogonal contrasts is one less than the number of levels of the independent variable.

All contrasts are evaluated by the same general procedure. First, the contrast is formalized as a set of contrast coefficients (also called *contrast weights*). Second, a specific $F$ ratio (denoted $F_\psi$) is computed. Finally, the probability associated with $F_\psi$ is evaluated. This last step changes with the type of analysis performed.

## Research Hypothesis as a Contrast Expression

When a research hypothesis is precise, it is possible to express it as a contrast. A research hypothesis, in general, can be expressed as a shape, a configuration, or a rank ordering of the experimental means. In all these cases, one can assign numbers that will reflect the predicted values of the experimental means. These numbers are called *contrast coefficients* when their mean is zero. To convert a set of numbers into a contrast, it suffices to subtract their mean from each of them. Often, for convenience, contrast coefficients are expressed with integers.

For example, assume that for a four-group design, a theory predicts that the first and second groups should be equivalent, the third group should perform better than these two groups, and the fourth group should do better than the third with an advantage of twice the gain of the third over the first and the second. When translated into a set of ranks, this prediction gives

| $C_1$ | $C_2$ | $C_3$ | $C_4$ | Mean |
|---|---|---|---|---|
| 1 | 1 | 2 | 4 | 2 |

After subtracting the mean, we get the following contrast:

| $C_1$ | $C_2$ | $C_3$ | $C_4$ | Mean |
|---|---|---|---|---|
| –1 | –1 | 0 | 2 | 0 |

In case of doubt, a good heuristic is to draw the predicted configuration of results, and then to represent the position of the means by ranks.

## A Priori Orthogonal Contrasts

### *For Multiple Tests*

When several contrasts are evaluated, several statistical tests are performed on the same data set, and this increases the probability of a Type I error (i.e., rejection of the null hypothesis when it is true). In order to control the Type I error at the level of the set (also known as the *family*) of contrasts, one needs to correct the $\alpha$ level used to evaluate each contrast. This correction for multiple contrasts can be done with the use of the Šidák equation, the Bonferroni (also known as Boole, or Dunn) inequality, or the Monte Carlo technique.

#### Šidák and Bonferroni

The probability of making *at least one* Type I error for a family of orthogonal (i.e., statistically independent) contrasts ($C$) is

$$\alpha[PF] = 1 - (1 - \alpha[PC])^{C}. \quad (1)$$

with $\alpha[PF]$ being the Type I error for the family of contrasts and $\alpha[PC]$ being the Type I error *per* contrast. This equation can be rewritten as

$$\alpha[PC] = 1 - (1 - \alpha[PF])^{1/C}. \quad (2)$$

This formula, called the Šidák equation, shows how to correct the $[PC]$ values used for each contrast.

Because the Šidák equation involves a fractional power, one can use an approximation known as the Bonferroni inequality, which relates $\alpha[PC]$ to $\alpha[PF]$:

$$\alpha[PC] \approx \frac{\alpha[PF]}{C}. \quad (3)$$

Šidák and Bonferroni are related by the inequality

$$\alpha[PC] = 1 - (1 - \alpha[PF])^{1/C} \geq \frac{\alpha[PF]}{C}. \quad (4)$$

They are, in general, very close to each other. As can be seen, the Bonferroni inequality is a pessimistic estimation. Consequently Šidák should be preferred. However, the Bonferroni inequality is more well known and hence is used and cited more often.

### Monte Carlo

The Monte Carlo technique can also be used to correct for multiple contrasts. The Monte Carlo technique consists of running a simulated experiment many times using random data, with the aim of obtaining a pattern of results showing what would happen just on the basis of chance. This approach can be used to quantify $\alpha[PF]$, the inflation of Type I error due to multiple testing. Equation 1 can then be used to set $\alpha[PC]$ in order to control the overall value of the Type I error.

As an illustration, suppose that six groups with 100 observations per group are created with data randomly sampled from a normal population. By construction, the $H_0$ is true (i.e., all population means are equal). Now, construct five *independent contrasts* from these six groups. For each contrast, compute an $F$ test. If the probability associated with the statistical index is smaller than $\alpha = .05$, the contrast is said to reach significance (i.e., $\alpha[PC]$ is used). Then have a computer redo the experiment 10,000 times. In sum, there are 10,000 experiments, 10,000 families of contrasts, and $5 \times 10{,}000 = 50{,}000$ contrasts. The results of this simulation are given in Table 1.

Table 1 shows that the $H_0$ is rejected for 2,403 contrasts out of the 50,000 contrasts actually performed (5 contrasts $\times$ 10,000 experiments). From these data, an estimation of $\alpha[PC]$ is computed as

**Table 1** Results of a Monte Carlo Simulation

| *Number of Families With* X *Type I Errors* | X: *Number of Type 1 Errors per Family* | *Number of Type I Errors* |
|---|---|---|
| 7,868 | 0 | 0 |
| 1,907 | 1 | 1,907 |
| 192 | 2 | 384 |
| 20 | 3 | 60 |
| 13 | 4 | 52 |
| 0 | 5 | 0 |
| 10,000 | | 2,403 |

*Notes:* Numbers of Type I errors when performing $C = 5$ contrasts for 10,000 analyses of variance performed on a six-group design when the $H_0$ is true. For example, 192 families out of the 10,000 have two Type I errors. This gives $2 \times 192 = 384$ Type I errors.

$$\alpha[PC] = \frac{\text{number of contrasts having reached significance}}{\text{total number of contrasts}} = \frac{2{,}403}{50{,}000} = .0479. \quad (5)$$

This value falls close to the theoretical value of $\alpha = .05$.

It can be seen also that for 7,868 experiments, no contrast reached significance. Correspondingly, for 2,132 experiments (10,000 – 7,868), at least one Type I error was made. From these data, $\alpha[PF]$ can be estimated as

$$\alpha[PF] = \frac{\text{number of families with at least 1 Type I error}}{\text{total number of families}} \quad 1.5 = \frac{2{,}132}{10{,}000} = .2132. \quad (6)$$

This value falls close to the theoretical value given by Equation 1:

$$\alpha[PF] = 1 - (1 - \alpha[PC])^C = 1 - (1 - .05)^5 = .226 .$$

### *Checking the Orthogonality of Two Contrasts*

Two contrasts are orthogonal (or independent) if their contrast coefficients are uncorrelated. Contrast coefficients have zero sum (and therefore a zero mean). Therefore, two contrasts, whose $A$ contrast coefficients are denoted $C_{a1}$ and $C_{a2}$, will be orthogonal if and only if

$$\sum_{a=1}^{A} C_{a,i} C_{a,j} = 0. \quad (7)$$

### *Computing Sum of Squares, Mean Square, and F*

The sum of squares for a contrast can be computed with the $C_a$ coefficients. Specifically, the sum of squares for a contrast is denoted $SS_\psi$ and is computed as

$$SS_\psi = \frac{S(\sum C_a M_{a.})^2}{\sum C_a^2} \quad (8)$$

where $S$ is the number of subjects in a group.

Also, because the sum of squares for a contrast has one degree of freedom, it is equal to the mean square of effect for this contrast:

**Table 2** Data From a Replication of an Experiment by Smith (1979)

| | *Experimental Context* | | | | |
|---|---|---|---|---|---|
| | *Group 1* | *Group 2* | *Group 3* | *Group 4* | *Group 5* |
| | *Same* | *Different* | *Imagery* | *Photo* | *Placebo* |
| | 25 | 11 | 14 | 25 | 8 |
| | 26 | 21 | 15 | 15 | 20 |
| | 17 | 9 | 29 | 23 | 10 |
| | 15 | 6 | 10 | 21 | 7 |
| | 14 | 7 | 12 | 18 | 15 |
| | 17 | 14 | 22 | 24 | 7 |
| | 14 | 12 | 14 | 14 | 1 |
| | 20 | 4 | 20 | 27 | 17 |
| | 11 | 7 | 22 | 12 | 11 |
| | 21 | 19 | 12 | 11 | 4 |
| $Y_a$ | 180 | 110 | 170 | 190 | 100 |
| $M_a$ | 18 | 11 | 17 | 19 | 10 |
| $M_a - M$ | 3 | – 4 | 2 | 4 | – 5 |
| $\sum(Y_{as} - M_a)^2$ | 218 | 284 | 324 | 300 | 314 |

*Source:* Adapted from Smith (1979).

*Note:* The dependent variable is the number of words recalled.

$$MS_\psi = \frac{SS_\psi}{df_\psi} = \frac{SS_\psi}{1} = SS_\psi. \tag{9}$$

The $F_\psi$ ratio for a contrast is now computed as

$$F_\psi = \frac{MS_\psi}{MS_{\text{error}}}\ . \tag{10}$$

### *Evaluating* F *for Orthogonal Contrasts*

Planned orthogonal contrasts are equivalent to independent questions asked of the data. Because of that independence, the current procedure is to act as if each contrast were the only contrast tested. This amounts to *not* using a correction for multiple tests. This procedure gives maximum power to the test. Practically, the null hypothesis for a contrast is tested by computing an $F$ ratio as indicated in Equation 10 and evaluating its $p$ value using a Fisher sampling distribution with $\nu_1 = 1$ and $\nu_2$ being the number of degrees of freedom of $MS_{\text{error}}$ [e.g., in independent measurement designs with $A$ groups and $S$ observations per group, $\nu_2 = A(S-1)$].

### *Example*

This example is inspired by an experiment by Steven M. Smith in 1979. The main purpose of this experiment was to show that one's being in the same mental context for learning and for testing leads to better performance than being in different contexts. During the learning phase, participants learned a list of 80 words in a room painted with an orange color, decorated with posters, paintings, and a decent amount of paraphernalia. A memory test was performed to give subjects the impression that the experiment was over. One day later, the participants were unexpectedly retested on their memory. An experimenter asked them to write down all the words from the list that they could remember. The test took place in five different experimental conditions. Fifty subjects (10 per group) were randomly assigned to one of the five experimental groups. The five experimental conditions were

1. *Same context*. Participants were tested in the same room in which they learned the list.
2. *Different context*. Participants were tested in a room very different from the one in which

**Table 3** ANOVA Table for a Replication of Smith's Experiment (1979)

| *Source* | *df* | *SS* | *MS* | *F* | *Pr(F)* |
|---|---|---|---|---|---|
| Experimental | 4 | 700.00 | 175.00 | 5.469** | .00119 |
| Error | 45 | 1,440.00 | 32.00 | | |
| Total | 49 | 2,1400.00 | | | |

*Source:* Adapted from Smith (1979).

*Note:* $^{**}p \leq .01$.

**Table 4** Orthogonal Contrasts for the Replication of Smith (1979)

| *Contrast* | *Group 1* | *Group 2* | *Group 3* | *Group 4* | *Group 5* | $\sum Ca$ |
|---|---|---|---|---|---|---|
| $\psi_1$ | +2 | −3 | +2 | +2 | −3 | 0 |
| $\psi_2$ | +2 | 0 | −1 | −1 | 0 | 0 |
| $\psi_3$ | 0 | 0 | +1 | −1 | 0 | 0 |
| $\psi_4$ | 0 | +1 | 0 | 0 | −1 | 0 |

*Source:* Adapted from Smith (1979).

they learned the list. The new room was located in a different part of the campus, painted grey, and looked very austere.

3. *Imaginary context.* Participants were tested in the same room as participants from Group 2. In addition, they were told to try to remember the room in which they learned the list. In order to help them, the experimenter asked them several questions about the room and the objects in it.
4. *Photographed context.* Participants were placed in the same condition as Group 3, and in addition, they were shown photos of the orange room in which they learned the list.
5. *Placebo context.* Participants were in the same condition as participants in Group 2. In addition, before starting to try to recall the words, they are asked to perform a warm-up task, namely, to try to remember their living room.

The data and ANOVA results of the replication of Smith's experiment are given in the Tables 2 and 3.

### Research Hypotheses for Contrast Analysis

Several research hypotheses can be tested with Smith's experiment. Suppose that the experiment was designed to test these hypotheses:

- *Research Hypothesis 1.* Groups for which the context at test matches the context during learning (i.e., is the same or is simulated by imaging or photography) will perform better than groups with different or placebo contexts.
- *Research Hypothesis 2.* The group with the same context will differ from the group with imaginary or photographed contexts.
- *Research Hypothesis 3.* The imaginary context group differs from the photographed context group.
- *Research Hypothesis 4.* The different context group differs from the placebo group.

#### *Contrasts*

The four research hypotheses are easily transformed into statistical hypotheses. For example, the first research hypothesis is equivalent to stating the following null hypothesis:

The means of the population for Groups 1, 3, and 4 have the same value as the means of the population for Groups 2 and 5.

This is equivalent to contrasting Groups 1, 3, and 4, on one hand, and Groups 2 and 5, on the other. This first contrast is denoted $\psi_1$:

$$\psi_1 = 2\mu_1 - 3\mu_2 + 2\mu_3 + 2\mu_4 + 3\mu_5.$$

The null hypothesis to be tested is

$$H_{0,1} : \psi_1 = 0.$$

The first contrast is equivalent to defining the following set of coefficients $C_a$:

| Gr.1 | Gr.2 | Gr.3 | Gr.4 | Gr.5 | $\sum_a C_a$ |
|---|---|---|---|---|---|
| +2 | −3 | +2 | +2 | −3 | 0 |

Note that the sum of the coefficients $C_a$ is zero, as it should be for a contrast. Table 4 shows all four contrasts.

### Are the Contrasts Orthogonal?

Now the problem is to decide whether the contrasts constitute an orthogonal family. We check that every pair of contrasts is orthogonal by using Equation 7. For example, Contrasts 1 and 2 are orthogonal because

**Table 5** Steps for the Computation of $SS_{\psi 1}$ of Smith (1979)

| *Group* | *Ma* | *Ca* | *CaMa* | $C_a^2$ |
|---|---|---|---|---|
| 1 | 18.00 | +2 | +36.00 | 4 |
| 2 | 11.00 | –3 | –33.00 | 9 |
| 3 | 17.00 | +2 | +34.00 | 4 |
| 4 | 19.00 | +2 | +38.00 | 4 |
| 5 | 10.00 | –3 | –30.00 | 9 |
| | | 0 | 45.00 | 30 |

*Source:* Adapted from Smith (1979).

$$\sum_{a=1}^{A=5} C_{a,1}C_{a,2} = (2 \times 2) + (-3 \times 0) + (2 \times -1) + (2 \times -1) + (-3 \times 0) + (0 \times 0) = 0.$$

#### *F* test

The sum of squares and $F_\psi$ for a contrast are computed from Equations 8 and 10. For example, the steps for the computations of $SS_{\psi 1}$ are given in Table 5.

$$\begin{aligned} SS_{\psi 1} &= \frac{S(\sum C_a M_{a.})^2}{\sum C_a^2} = \frac{10 \times (45.00)^2}{30} = 675.00 \\ MS_{\psi 1} &= 675.00 \\ F_{\psi 1} &= \frac{MS_{\psi 1}}{MS_{error}} = \frac{675.00}{32.00} = 21.094 . \end{aligned} \tag{11}$$

The significance of a contrast is evaluated with a Fisher distribution with 1 and $A(S-1) = 45$ degrees of freedom, which gives a critical value of 4.06 for $\alpha = .05$ (7.23 for $\alpha = .01$). The sums of squares for the remaining contrasts are $SS_{\psi.2} = 0$, $SS_{\psi.3} = 20$, and $SS_{\psi.4} = 5$ with 1 and $A(S-1) = 45$ degrees of freedom. Therefore, $\psi_2$, $\psi_3$, and $\psi_4$ are nonsignificant. Note that the sums of squares of the contrasts add up to $SS_{\text{experimental}}$. That is,

$$\begin{aligned} SS_{\text{experimental}} &= SS_{\psi.1} + SS_{\psi.2} + SS_{\psi.3} + SS_{\psi.4} \\ &= 675.00 + 0.00 + 20.00 + 5.00 \\ &= 700.00 . \end{aligned}$$

When the sums of squares are orthogonal, the degrees of freedom are added the same way as the sums of squares are. This explains why the maximum number of orthogonal contrasts is equal to the number of degrees of freedom of the experimental sum of squares.

### *A Priori* Nonorthogonal Contrasts

So orthogonal contrasts are relatively straightforward because each contrast can be evaluated on its own. Nonorthogonal contrasts, however, are more complex. The main problem is to assess the importance of a given contrast conjointly with the other contrasts. There are currently two (main) approaches to this problem. The classical approach corrects for multiple statistical tests (e.g., using a Šidák or Bonferroni correction), but essentially evaluates each contrast as if it were coming from a set of orthogonal contrasts. The multiple regression (or modern) approach evaluates each contrast as a predictor from a set of nonorthogonal predictors and estimates its *specific* contribution to the explanation of the dependent variable. The classical approach evaluates each contrast for itself, whereas the multiple regression approach evaluates each contrast as a member of a set of contrasts and estimates the specific contribution of each contrast in this set. For an orthogonal set of contrasts, the two approaches are equivalent.

### The Classical Approach

Some problems are created by the use of multiple nonorthogonal contrasts. The most important one is that the greater the number of contrasts, the greater the risk of a Type I error. The general strategy adopted by the classical approach to this problem is to correct for multiple testing.

#### Šidák and Bonferroni Corrections

When a family's contrasts are nonorthogonal, Equation 10 gives a lower bound for $[PC]$. So, instead of having the equality, the following inequality, called the Šidák inequality, holds:

$$\alpha[PF] \leq 1 - (1 - \alpha[PC])^C. \tag{12}$$

This inequality gives an upper bound for $\alpha[PF]$, and therefore the real value of $\alpha[PF]$ is smaller than its estimated value.

As earlier, we can approximate the Šidák inequality by Bonferroni as

$$\alpha[PF] < C\alpha[PC]. \tag{13}$$

**Table 6** Nonorthogonal Contrasts for the Replication of Smith (1979)

| Contrast | Group 1 | Group 2 | Group 3 | Group 4 | Group 5 | $\sum C_a$ |
|---|---|---|---|---|---|---|
| $\psi_1$ | 2 | –3 | 2 | 2 | –3 | 0 |
| $\psi_2$ | 3 | 3 | –2 | –2 | –2 | 0 |
| $\psi_3$ | 1 | –4 | 1 | 1 | 1 | 0 |

*Source:* Adapted from Smith (1979).

And, as earlier, Šidák and Bonferroni are linked to each other by the inequality

$$\alpha[PF] \leq 1 - (1 - \alpha[PC])^C < C\alpha[PC]. \qquad (14)$$

**Example**

Let us go back to Smith's (1979) study (see Table 2). Suppose that Smith wanted to test these three hypotheses:

- *Research Hypothesis 1.* Groups for which the context at test matches the context during learning will perform better than groups with different contexts;
- *Research Hypothesis 2.* Groups with real contexts will perform better than those with imagined contexts;
- *Research Hypothesis 3.* Groups with any context will perform better than those with no context.

These hypotheses can easily be transformed into the set of contrasts given in Table 6. The values of $F_\psi$ were computed with Equation 10 (see also Table 3) and are in shown in Table 7, along with their $p$ values. If we adopt a value of $\alpha[PF] = .05$, a Šidák correction will entail evaluating each contrast at the $\alpha$ level of $\alpha[PF] = .0170$ (Bonferroni will give the approximate value of $\alpha[PF] = .0167$). So, with a correction for multiple comparisons one can conclude that Contrasts 1 and 3 are significant.

### *Multiple Regression Approach*

ANOVA and multiple regression are equivalent if one uses as many predictors for the multiple regression analysis as the number of degrees of freedom of the independent variable. An obvious choice for the predictors is to use a set of contrast coefficients. Doing so makes contrast analysis a particular case of multiple regression analysis.

**Table 7** $F_\psi$ Values for the Nonorthogonal Contrasts From the Replication of Smith (1979)

| Contrast | $r_{Y.\psi}$ | $r^2_{Y.\psi}$ | $F_\psi$ | $pF_\psi$ |
|---|---|---|---|---|
| $\psi_1$ | .9820 | .9643 | 21.0937 | $< .0001$ |
| $\psi_2$ | –.1091 | .0119 | 0.2604 | .6123 |
| $\psi_3$ | .5345 | .2857 | 6.2500 | .0161 |

*Source:* Adapted from Smith (1979).

When used with a set of orthogonal contrasts, the multiple regression approach gives the same results as the ANOVA-based approach previously described. When used with a set of *nonorthogonal* contrasts, multiple regression quantifies the *specific* contribution of each contrast as the semipartial coefficient of correlation between the contrast coefficients and the dependent variable. The multiple regression approach can be used for nonorthogonal contrasts as long as the following constraints are satisfied:

1. There are no more contrasts than the number of degrees of freedom of the independent variable.
2. The set of contrasts is linearly independent (i.e., not multicollinear). That is, no contrast can be obtained by combining the other contrasts.

**Example**

Let us go back once again to Smith's (1979) study of learning and recall contexts. Suppose we take our three contrasts (see Table 6) and use them as predictors with a standard multiple regression program. We will find the following values for the semipartial correlation between the contrasts and the dependent variable:

$$\psi_1 : r^2_{Y.C_{a,1}|C_{a,2}C_{a,3}} = .1994$$
$$\psi_2 : r^2_{Y.C_{a,2}|C_{a,1}C_{a,3}} = .0000$$
$$\psi_3 : r^2_{Y.C_{a,3}|C_{a,1}C_{a,2}} = .0013,$$

with $r^2_{Y.C_{a,1}|C_{a,2}C_{a,3}}$ being the squared correlation of $\psi_1$ and the dependent variable with the effects of $\psi_2$ and $\psi_3$ partialled out. To evaluate the significance of each contrast, we compute an $F$ ratio for the corresponding semipartial coefficients of correlation. This is done with the following formula:

$$F_{Y.C_{a,i}|C_{a,k}C_{a,\ell}} = \frac{r^2_{Y.C_{a,i}|C_{a,k}C_{a,\ell}}}{1 - r^2_{Y.A}} \times df_{\text{residual}}. \quad (15)$$

This results in the following $F$ ratios for the Smith example:

$$\psi_1 : F_{Y.C_{a,1}|C_{a,2}C_{a,3}} = 13.3333, \quad p = .0007;$$
$$\psi_2 : F_{Y.C_{a,2}|C_{a,1}C_{a,3}} = 0.0000, \quad p = 1.0000;$$
$$\psi_3 : F_{Y.C_{a,3}|C_{a,1}C_{a,2}} = 0.0893, \quad p = .7665.$$

These $F$ ratios follow a Fisher distribution with $\nu_1 = 1$ and $\nu_2 = 45$ degrees of freedom. $F_{\text{critical}} = 4.06$ when $\alpha = .05$. In this case, $\psi_1$ is the only contrast reaching significance (i.e., with $F_\psi > F_{\text{critical}}$). The comparison with the classic approach shows the drastic differences between the two approaches.

## A Posteriori Contrasts

For a posteriori contrasts, the family of contrasts is composed of all the possible contrasts even if they are not explicitly made. Indeed, because one chooses one of the contrasts to be made a posteriori, this implies that one has *implicitly* made and judged uninteresting *all* the possible contrasts that have not been made. Hence, whatever the number of contrasts actually performed, the family is composed of all the possible contrasts. This number grows very fast: A conservative estimate indicates that the number of contrasts that can be made on $A$ groups is equal to

$$1 + \{[(3^A - 1)/2] - 2^A\}. \quad (16)$$

So using a Šidàk or Bonferroni approach will not have enough power to be useful.

### *Scheffé's Test*

Scheffé's test was devised to test all possible contrasts a posteriori while maintaining the overall Type I error level for the family at a reasonable level, as well as trying to have a conservative but relatively powerful test. The general principle is to ensure that no discrepant statistical decision can occur. A discrepant decision would occur if the omnibus test would fail to reject the null hypothesis, but one a posteriori contrast could be declared significant.

In order to avoid such a discrepant decision, the Scheffé approach first tests any contrast as if it were the largest possible contrast whose sum of squares is equal to the experimental sum of squares (this contrast is obtained when the contrast coefficients are equal to the deviations of the group means to their grand mean) and, second, makes the test of the largest contrast equivalent to the ANOVA omnibus test. So if we denote by $F_{\text{critical, omnibus}}$ the critical value for the ANOVA omnibus test (performed on $A$ groups), the largest contrast is equivalent to the omnibus test if its $F_\psi$ is tested against a critical value equal to

$$F_{\text{critical, Scheffé}} = (A - 1) \times F_{\text{critical, omnibus}}. \quad (17)$$

Equivalently, $F_\psi$ can be divided by $(A - 1)$, and its probability can be evaluated with a Fisher distribution with $\nu_1 = (A - 1)$ and $\nu_2$ being equal to the number of degrees of freedom of the mean square error. Doing so makes it *impossible* to reach a discrepant decision.

#### 4.1.1 An example: Scheffé

Suppose that the $F_\psi$ ratios for the contrasts computed in Table 7 were obtained a posteriori. The critical value for the ANOVA is obtained from a Fisher distribution with $\nu_1 = A - 1 = 4$ and $\nu_2 = A(S - 1) = 45$. For $\alpha = .05$, this value is equal to $F_{\text{critical, omnibus}} = 2.58$. In order to evaluate whether any of these contrasts reaches significance, one needs to compare them to the critical value of

$$\begin{aligned} F_{\text{critical, Scheffé}} &= (A - 1) \times F_{\text{criticalm, omnibus}}. \\ &= 4 \times 2.58 = 10.32. \end{aligned} \quad (18)$$

With this approach, only the first contrast is considered significant.

*Hervé Abdi and Lynne J. Williams*

***See also*** Analysis of Variance (ANOVA); Type I Error

**Further Readings**

Abdi, H., Edelman, B., Valentin, D., & Dowling, W. J. (2009). *Experimental design and analysis for psychology.* Oxford, UK: Oxford University Press.

Rosenthal, R., & Rosnow, R. L. (2003). *Contrasts and effect sizes in behavioral research: A correlational approach*. Boston: Cambridge University Press.

Smith, S. M. (1979). Remembering in and out of context. *Journal of Experimental Psychology: Human Learning and Memory, 5,* 460–471.

# Control Group

In experimental research, it is important to confirm that results of a study are actually due to an independent or manipulated variable rather than to other, extraneous variables. In the simplest case, a research study contrasts two groups, and the independent variable is present in one group but not the other. For example, in a health research study, one group may receive a medical treatment, and the other does not. The first group, in which treatment occurs, is called the *experimental group,* and the second group, in which treatment is withheld, is called the *control group*. Therefore, when experimental studies use control and experimental groups, ideally the groups are equal on all factors except the independent variable. The purpose, then, of a control group is to provide a comparative standard in order to determine whether an effect has taken place in the experimental group. As such, the term *control group* is sometimes used interchangeably with the term *baseline group* or *contrast group*. The following discussion outlines key issues and several varieties of control groups employed in experimental research.

## Random Assignment

In true experiments, subjects are assigned randomly to either a control or an experimental group. If the groups are alike in all ways except for the treatment administered, then the effects of that treatment can be tested without ambiguity. Although no two groups are exactly alike, random assignment, especially with large numbers, evens differences out. By and large, randomized group assignment results in groups that are equivalent, with the exception of the independent variable, or treatment of interest.

## Placebo Control Groups

A *placebo* is a substance that appears to have an effect but is actually inert. When individuals are part of a placebo control group, they believe that they are receiving an effective treatment, when it is in fact a placebo. An example of a placebo study might be found in medical research in which researchers are interested in the effects of a new medication for cancer patients. The experimental group would receive the treatment under investigation, and the control group might receive an inert substance. In a double-blind placebo study, neither the participants nor the experimenters know who received the placebo until the observations are complete. Double-blind procedures are used to prevent experimenter expectations or experimenter bias from influencing observations and measurements.

Another type of control group is called a waiting list control. This type of control group is often used to assess the effectiveness of a treatment. In this design, all participants may experience an independent variable, but not at the same time. The experimental group receives a treatment and is then contrasted, in terms of the effects of the treatment, with a group awaiting treatment. For example, the effects of a new treatment for depression may be assessed by comparing a treated group, that is, the experimental group, with individuals who are on a wait list for treatment. Individuals on the wait list control may be treated subsequently.

When participants in an experimental group experience various types of events or participate for varying times in a study, the control group is called a *yoked control group*. Each participant of the control group is "yoked" to a member of the experimental group. As an illustration, suppose a study is interested in assessing the effects of students' setting their own learning goals. Participants in this study might be yoked on the instructional time that they receive on a computer. Each participant in the yoked control (no goal setting) would be yoked to a corresponding student in the experimental group on the amount of time he or she spent on learning from the computer. In this way,

the amount of instruction experienced is held constant between the groups.

*Matching procedures* are similar to yoking, but the matching occurs on characteristics of the participant, not the experience of the participant during the study. In other words, in matching, participants in the control group are matched with participants in the experimental group so that the two groups have similar backgrounds. Participants are often matched on variables such as age, gender, and socioeconomic status. Both yoked and matched control groups are used so that participants are as similar as possible.

### Ethics in Control Groups

The decision regarding who is assigned to a control group has been the topic of ethical concerns. The use of placebo controlled trials is particularly controversial because many ethical principles must be considered. The use of a wait-list control group is generally viewed as a more ethical approach when examining treatment effects with clinical populations. If randomized control procedures were used, a potentially effective treatment would have to be withheld. The harm incurred from not treating patients in a control group is usually determined, on balance, to be greater than the added scientific rigor that random assignment might offer.

Developed by the World Medical Association, the Declaration of Helsinki is an international document that provides the groundwork for general human research ethics. Generally, the guiding principle when one is using control groups is to apply strict ethical standards so that participants are not at risk of harm. Overall, the ethics involved in the use of control groups depends on the context of the scientific question.

*Jennie K. Gill and John Walsh*

***See also*** Double-Blind Procedure; Experimental Design; Internal Validity; Sampling

#### Further Readings

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis for field settings*. Boston: Houghton Mifflin.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

World Medical Association. (2002). Declaration of Helsinki: Ethical principles for medical research involving human subjects. *Journal of Postgraduate Medicine, 48*, 206–208.

## Control Variables

In experimental and observational design and data analysis, the term *control variable* refers to variables that are not of primary interest (i.e., neither the exposure nor the outcome of interest) and thus constitute an extraneous or third factor whose influence is to be controlled or eliminated. The term refers to the investigator's desire to estimate an effect (such as a measure of association) of interest that is independent of the influence of the extraneous variable and free from bias arising from differences between exposure groups in that third variable.

Extraneous variables of this class are usually those variables described as potential confounders in some disciplines. Controlling for a potential confounder, which is not an effect modifier or mediator, is intended to isolate the effect of the exposure of interest on the outcome of interest while reducing or eliminating potential bias presented by differences in the outcomes observed between exposed and unexposed individuals that are attributable to the potential confounder. Control is achieved when the potential confounder cannot vary between the exposure groups, and thus the observed relationship between the exposure and outcome of interest is independent of the potential confounder.

As an example, if an investigator is interested in studying the rate of a chemical reaction (the outcome) and how it differs with different reagents (the exposures), the investigator may choose to keep the temperature of each reaction constant among the different reagents being studied so that temperature differences could not affect the outcomes.

Potential control variables that are mediators in another association of interest, as well as potential

control variables that are involved in a statistical reaction with other variables in the study, are special cases which must be considered separately. This entry discusses the use of control variables during the design and analysis stages of a study.

## Design Stage

There are several options for the use of control variables at the design stage. In the example about rates of reaction mentioned earlier, the intention was to draw conclusions, at the end of the series of experiments, regarding the relationship between the reaction rates and the various reagents. If the investigator did not keep the temperature constant among the series of experiments, difference in the rate of reaction found at the conclusion of the study may have had nothing to do with different reagents, but be solely due to differences in temperature or some combination of reagent and temperature. Restricting or specifying a narrow range of values for one or more potential confounders is frequently done in the design stage of the study, taking into consideration several factors, including ease of implementation, convenience, simplified analysis, and expense. A limitation on restriction may be an inability to infer the relationship between the restricted potential confounder and the outcome and exposure. In addition, residual bias may occur, owing to incomplete control (referred to as *residual confounding*).

Matching is a concept related to restriction. Matching is the process of making the study group and control group similar with regard to potential confounds. Several different methods can be employed, including *frequency matching, category matching, individual matching,* and *caliper matching*. As with restriction, the limitations of matching include the inability to draw inferences about the control variable(s). Feasibility can be an issue, given that a large pool of subjects may be required to find matches. In addition, the potential for residual confounding exists.

Both matching and restriction can be applied in the same study design for different control variables.

## The Analysis Stage

There are several options for the use of control variables at the analysis stage. Separate analysis can be undertaken for each level of a potential confounder. Within each unique value (or homogeneous stratum) of the potential confounder, the relationship of interest may be observed that is not influenced by differences between exposed and unexposed individuals attributable to the potential confounder. This technique is another example of restriction.

Estimates of the relationship of interest independent of the potential confounder can also be achieved by the use of a matched or stratified approach in the analysis. The estimate of interest is calculated at all levels (or several theoretically homogeneous or equivalent strata) of the potential confounder, and a weighted, average effect across strata is estimated. Techniques of this kind include the *Mantel–Haenszel stratified analysis*, as well as stratified (also called matched or conditional) regression analyses. These approaches typically assume that the stratum-specific effects are not different (i.e., no effect modification or statistical interaction is present). Limitations of this method are related to the various ways strata can be formed for the various potential confounders, and one may end up with small sample sizes in many strata, and therefore the analysis may not produce a reliable result.

The most common analytic methods for using control variables is analysis of covariance and multiple generalized linear regression modeling. Regression techniques estimate the relationship of interest conditional on a fixed value of the potential confounder, which is analogous to holding the value of the potential confounder constant at the level of third variable. By default, model parameters (intercept and beta coefficients) are interpreted as though potential confounders were held constant at their zero values. Multivariable regression is relatively efficient at handling small numbers and easily combines variables measured on different scales.

Where the potential control variable in question is involved as part of a statistical interaction with an exposure variable of interest, holding the control variable constant at a single level through restriction (in either the design or analysis) will allow estimation of the effect of the exposure of interest and the outcome that is independent of the third variable, but the effect measured applies only to (or is conditional on) the selected level of the

potential confounder. This would also be the stratum-specific or conditional effect. For example, restriction of an experiment to one gender would give the investigator a gender-specific estimate of effect.

If the third variable in question is part of a true interaction, the other forms of control, which permit multiple levels of the third variable to remain in the study (e.g., through matching, statistical stratification, or multiple regression analysis), should be considered critically before being applied. Each of these approaches ignores the interaction and may serve to mask its presence.

*Jason D. Pole and Susan J. Bondy*

***See also*** Bias; Confounding; Interaction; Matching

**Further Readings**

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton Mifflin.

## Convenience Sampling

Few terms or concepts in the study of research design are as self-explanatory as convenience sampling. Convenience sampling (sometimes called *accidental sampling*) is the selection of a sample of participants from a population based on how convenient and readily available that group of participants is. It is a type of nonprobability sampling that focuses on a sample that is easy to access and readily available. For example, if one were interested in knowing the attitudes of a group of 1st-year college students toward binge drinking, a convenience sample would be those students enrolled in an introductory biology class.

The advantages of convenience sampling are clear. Such samples are easy to obtain, and the cost of obtaining them is relatively low. The disadvantages of convenience sampling should be equally clear. Results from studies using convenience sampling are not very generalizable to other settings, given the narrow focus of the technique. For example, using those biology 101 students would not only limit the sample to 1st-year students in that class, but also to those students enrolled in biology and all their characteristics. However, in spite of any shortcomings, convenience sampling is still an effective tool to use in pilot settings, when instruments may still be under development and interventions are yet to be fully designed and approved.

*Neil J. Salkind*

***See also*** Cluster Sampling; Experience Sampling Method; Nonprobability Sampling; Probability Sampling; Proportional Sampling; Quota Sampling; Random Sampling; Sampling; Sampling Error; Stratified Sampling; Systematic Sampling

**Further Readings**

Hultsch, D. F., MacDonald, S. W. S., Hunter, M. A., Maitland, S. B., & Dixon, R. A. (2002). Sampling and generalisability in developmental research: Comparison of random and convenience samples of older adults. *International Journal of Behavioral Development, 26*(4), 345–359.

Burke, T. W., Jordan, M. L., & Owen, S. S. (2002). A cross-national comparison of gay and lesbian domestic violence. *Journal of Contemporary Criminal Justice, 18*(3), 231–257.

## "Convergent and Discriminant Validation by the Multitrait–Multimethod Matrix"

Psychology as an empirical science depends on the availability of valid measures of a construct. Validity means that a measure (e.g., a test or questionnaire) adequately assesses the construct (trait) it intends to measure. In their 1959 article "Convergent and Discriminant Validation by the Multitrait–Multimethod Matrix," Donald T. Campbell and Donald W. Fiske proposed a way of test validation based on the idea that it is not sufficient to consider a single operationalization of a construct but that multiple measures are necessary. In order to validate a measure, scientists first have to define the construct to be measured in a literary form and then must generate at least two measures that are as different as possible but that are each adequate for

measuring the construct (*multiple operationalism* in contrast to *single operationalism*). These two measures should strongly correlate but differ from measures that were created to assess different traits.

Campbell and Fiske distinguished between four aspects of the validation process that can be analyzed by means of the multitrait–multimethod (MTMM) matrix. First, convergent validity is proven by the correlation of independent measurement procedures for measuring the same trait. Second, new measures of a trait should show low correlations with measures of other traits from which they should differ (discriminant validity). Third, each test is a trait–method unit. Consequently, interindividual differences in test scores can be due to measurement features, as well as to the content of the trait. Fourth, in order to separate method- from trait-specific influences, and to analyze discriminant validity, more than one trait and more than one method have to be considered in the validation process.

## Convergent Validity

Convergent validity evidence is obtained if the correlations of independent measures of the same trait (monotrait–heteromethod correlations) are significantly different from 0 and sufficiently large. Convergent validity differs from reliability in the type of methods considered. Whereas reliability is proven by correlations of maximally similar methods of a trait (monotrait–monomethod correlations), the proof of convergent validity is the stronger, the more independent the methods are. For example, reliability of a self-report extroversion questionnaire can be analyzed by the correlations of two test halves of this questionnaire (*split-half reliability*) whereas the convergent validity of the questionnaire can be scrutinized by its correlation with a peer report of extroversion. According to Campbell and Fiske, independence of methods is a matter of degree, and they consider reliability and validity as points on a continuum from reliability to validity. Heterotrait–monomethod correlations that do not significantly differ from 0 or are relatively low could indicate that one of the two measures or even both measures do not appropriately measure the trait (low convergent validity). However, a low correlation could also show that the two different measures assess different components of a trait that are functionally different. For example, a low correlation between an observational measure of anger and the self-reported feeling component of anger could indicate individuals who regulated their visible anger expression. In this case, a low correlation would not indicate that the self-report is an invalid measure of the feeling component and that the observational measure is an invalid indicator of overt anger expression. Instead, the two measures could be valid measures of the two different components of the anger episode that they are intended to measure, and different methods may be necessary to appropriately assess these different components. It is also recommended that one consider traits that are as independent as possible. If two traits are considered independent, the heterotrait–monomethod correlations should be 0. Differences from 0 indicate the degree of a common method effect.

## Discriminant Validity

Discriminant validity evidence is obtained if the correlations of variables measuring different traits are low. If two traits are considered independent, the correlations of the measures of these traits should be 0. Discriminant validity requires that both the heterotrait–heteromethod correlations (e.g., correlation of a self-report measure of extroversion and a peer report measure of neuroticism) and the heterotrait–monomethod correlations (e.g., correlations between self-report measures of extroversion and neuroticism) be small. These heterotrait correlations should also be smaller than the monotrait–heteromethod correlations (e.g., self- and peer report correlations of extroversion) that indicate convergent validity. Moreover, the patterns of correlations should be similar for the monomethod and the heteromethod correlations of different traits.

## Impact

According to Robert J. Sternberg, Campbell and Fiske's article is the most often cited paper that has ever been published in *Psychological Bulletin* and is one of the most influential publications in psychology. In an overview of then-available MTMM matrices, Campbell and Fiske concluded that

almost none of these matrices fulfilled the criteria they described. Campbell and Fiske considered the validation process as an iterative process that leads to better methods for measuring psychological constructs, and they hoped that their criteria would contribute to the development of better methods. However, 33 years later, in 1992, they concluded that the published MTMM matrices were still unsatisfactory and that many theoretical and methodological questions remained unsolved. Nevertheless, their article has had an enormous influence on the development of more advanced statistical methods for analyzing the MTMM matrix, as well as for the refinement of the validation process in many areas of psychology.

*Michael Eid*

***See also*** Construct Validity; MBESS; Multitrait–Multimethod Matrix; Triangulation; Validity of Measurement

**Further Readings**

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Eid, M., & Diener, E. (2006a). *Handbook of multimethod measurement in psychology.* Washington, DC: American Psychological Association.

Fiske, D. W. & Campbell, D. T. (1992). Citations do not solve problems. *Psychological Bulletin, 112,* 393–395.

Shrout, P. E., & Fiske, S. T. (Eds.). (1995). *Personality research, methods, and theory: A festschrift honoring Donald W. Fiske*. Hillsdale, NJ: Lawrence Erlbaum.

Sternberg, R. J. (1992). *Psychological Bulletin*'s top 10 "hit parade." *Psychological Bulletin, 112,* 387–388.

# Copula Functions

The word *copula* is a Latin noun that means a link and is used in grammar to describe the part of a proposition that connects the subject and predicate. Abe Sklar in 1959 was the first to introduce the word *copula* in a mathematical or statistical sense in a theorem describing the functions that join together one-dimensional distribution functions to form multivariate distribution functions. He called this class of functions *copulas.* In statistics, a copula is a function that links an $n$-dimensional cumulative distribution function to its one-dimensional margins and is itself a continuous distribution function characterizing the dependence structure of the model.

Recently, in multivariate modeling, much attention has been paid to copulas or *copula functions.* It can be shown that outside the elliptical world, correlation cannot be used to characterize the dependence between two series. To say it differently, the knowledge of two marginal distributions and the correlation does not determine the bivariate distribution of the underlying series. In this context, the only dependence function able to summarize all the information about the comovements of the two series is a copula function. Indeed, a multivariate distribution is fully and uniquely characterized by its marginal distributions and its dependence structure as represented by the copula.

## Definition and Properties

In what follows, the definition of a copula is provided in the bivariate case.

A copula is a function $C{:}[0,1]\times[0,1]\rightarrow[0,1]$ with the following properties:

1. For every $u, v$ in $[0,1]$, $C(u,0) = 0 = C(0,v)$, and $C(u,1) = u$ *and*$C(1,v) = v$.
2. For every $u_1, u_2, v_1, v_2$ in $[0,1]$, such that $u_1 \le u_2$ and $v_1 \le v_2$,$C(u_2,v_2) - C(u_2,v_1) - C(u_1,v_2) + C(u_1,v_1) \ge 0$.

An example is the product copula $C^{\perp}(u,v) = uv$, which is a very important copula because it characterizes independent random variables when the distribution functions are continuous.

One important property of copulas is the *Fréchet–Hoeffding bounds inequality,* given by

$$W(u,v) \le C(u,v) \le M(u,v),$$

where $W$ and $M$ are also copulas referred to as *Fréchet–Hoeffding* lower and upper bounds, respectively, and defined by $W(u,v) = \max(u + v - 1, 0)$ and $M(u,v) = \min(u,v)$.

Much of the usefulness of copulas in nonparametric statistics is due to the fact that for strictly monotone transformations of the random variables under interest, copulas are either invariant or

change in predictable ways. Specifically, let $X$ and $Y$ be continuous random variables with copula $C_{X,Y}$, and let $f$ and $g$ be strictly monotone functions on Ran$X$ and Ran$Y$ (Ran: Range), respectively.

1. If $f$ and $g$ are strictly increasing, then $C_{f(X),g(Y)}(u,v) = C_{X,Y}(u,v)$.
2. If $f$ is strictly increasing and $g$ is strictly decreasing, then $C_{f(X),g(Y)}(u,v) = u - C_{X,Y}(u,1-v)$.
3. If $f$ is strictly decreasing and $g$ is strictly increasing, then $C_{f(X),g(Y)}(u,v) = v - C_{X,Y}(1-u,v)$.
4. If $f$ and $g$ are strictly decreasing, then $C_{f(X),g(Y)}(u,v) = u + v - 1 + C_{X,Y}(1-u,1-v)$.

## Sklar's Theorem for Continuous Distributions

Sklar's theorem defines the role that copulas play in the relationship between multivariate distribution functions and their univariate margins.

### *Sklar's Theorem in the Bivariate Case*

Let $H$ be a joint distribution function with continuous margins $F$ and $G$. Then there exists a *unique* copula $C$ such that for all $x, y \in \overline{R}$,

$$H(x,y) = C(F(x), G(y)). \tag{1}$$

Conversely, if $C$ is a copula and $F$ and $G$ are distribution functions, then the function $H$ defined by Equation 1 is a joint distribution function with margins $F$ and $G$.

A multivariate version of this theorem exists and is presented hereafter.

### *Sklar's Theorem in* n *Dimensions*

For any multivariate distribution function $F(x_1, x_2, \ldots, x_n) = P(X_1 \le x_1, X_2 \le, x_2, \ldots, X_n \le x_n)$ with continuous marginal functions $F_i(x_i) = P(X_i \le x_i)$ for $1 \le i \le n$, there exists a *unique* function $C(u_1, u_2, \ldots, u_n)$, called the copula and defined on $[0,1]^n \to [0,1]$, such that for all $(x_1, x_2, \ldots, x_n) \in R^n$,

$$F(x_1, x_2, \ldots, x_n) = C(F_1(x_1), F_2(x_2), \ldots, F_n(x_n)). \tag{2}$$

Conversely, if $C(u_1, u_2, \ldots, u_n)$ is a copula and $F_i(x_i)$, $i = 1, 2, \ldots, n$, are the distribution functions of $n$ random variables $X_i$, then $F(x_1, x_2, \ldots, x_n)$ defined above is an $n$-variate distribution function with margins $F_i, i = 1, 2, \ldots, n$.

The expression of the copula function in Equation 2 is given in terms of cumulative distribution functions. If we further assume that each $F_i$ and $C$ is differentiable, the joint density $f(x_1, x_2, \ldots, x_n)$ corresponding to the cumulative distribution $F(x_1, x_2, \ldots, x_n)$ can easily be obtained using

$$f(x_1, x_2, \ldots, x_n) = \frac{\partial^n F(x_1, x_2, \ldots, x_n)}{\partial x_1 \, \partial x_2 \ldots \partial x_n}.$$

This gives Equation 3:

$$f(x_1, x_2, \ldots, x_n) = f_1(x_1) \times f_2(x_2) \times \ldots \times f_n(x_n) \times c(u_1, u_2, \ldots, u_n) \tag{3}$$

where the $f_i$s are the marginal densities corresponding to the $F_i$s, the $u_i$s are defined as $u_i = F_i(x_i)$, and $c(u_1, u_2, \ldots, u_n)$ is the density of the copula $C$ obtained as follows:

$$c(u_1, u_2, \ldots, u_n) = \frac{\partial^n C(u_1, u_2, \ldots, u_n)}{\partial u_1 \, \partial u_2 \ldots \partial u_n}. \tag{4}$$

In contrast to the traditional modeling approach that decomposes the joint density as a product of marginal and conditional densities, Equation 3 states that, under appropriate conditions, the joint density can be written as a product of the marginal densities and the copula density. From Equation 3, it is clear that the density $c(u_1, u_2, \ldots, u_n)$ encodes information about the dependence structure among the $X_i$s, and the $f_i$s describe the marginal behaviors. It thus shows that copulas represent a way to extract the dependence structure from the joint distribution and to extricate the dependence and marginal behaviors. Hence, copula functions offer more flexibility in modeling multivariate random variables. This flexibility contrasts with the traditional use of the multivariate normal distribution, in which the margins are assumed to be Gaussian and linked through a linear correlation matrix.

## Survival Copulas

A survival function $\overline{F}$ is defined as $\overline{F}(x) = P[X > x] = 1 - F(x)$, where $F$ denotes the distribution function of $X$. For a pair $(X, Y)$ of random variables with joint distribution function $H$, the joint survival function is given by $\overline{H}(x, y) = P[X > x\ ,\ Y > y]$ with margins $\overline{F}$ and $\overline{G}$, which are the univariate survival functions. Let $C$ be the copula of $X$ and $Y$; then

$$\begin{aligned}\overline{H}(x,y) &= 1 - F(x) - G(y) + H(x,y)\\ &= \overline{F}(x) + \overline{G}(y) - 1 + C(F(x), G(y))\\ &= \overline{F}(x) + \overline{G}(y) - 1 + C\big(1 - \overline{F}(x), 1 - \overline{G}(y)\big)\end{aligned}.$$

Let $\hat{C}$ be defined from $[0,1] \times [0,1]$ into $[0,1]$ by $\hat{C}(u, v) = u + v - 1 + C(1 - u, 1 - v)$, then $\overline{H}(x, y) = \hat{C}\big(\overline{F}(x), \overline{G}(y)\big)$.

Note that $\hat{C}$ links the joint survival function to its univariate margins in the same way a copula joins a bivariate cumulative function to its margins. Hence, $\hat{C}$ is a copula and is referred to as the survival copula of $X$ and $Y$. Note also that the survival copula $\hat{C}$ is different from the joint survival function $\overline{C}$ for two uniform random variables whose joint distribution function is the copula $C$.

## Simulation

Copulas can be used to generate a sample from a specified joint distribution. Such samples can then be used to study mathematical models of real-world systems or for statistical studies.

Various well-known procedures are used to generate independent uniform variates and to obtain samples from a given univariate distribution. More specifically, to obtain an observation $x$ of a random variable $X$ with distribution function $F$, the following method, called the *inverse distribution function method,* can be used:

- Generate a uniform variate $u$
- Set $x = F^{-1}(u)$, where $F^{-1}$is any quasi-inverse of $F$.

By virtue of Sklar's theorem, we need only to generate a pair $(u, v)$ of observations of uniform random variables $(U, V)$ whose distribution function is $C$, the copula of $X$ and $Y$, and then transform those uniform variates via the inverse distribution function method. One procedure for generating such a pair $(u, v)$ of uniform variates is the conditional distribution method. For this method, we need the conditional distribution function for $V$ given $U = u$, denoted $C_u(v)$ and defined as follows:

$$\begin{aligned}C_u(v) &= P[V \le v | U = u]\\ &= \lim_{\Delta u \to 0} \frac{C(u + \Delta u\ ,\ v) - C(u, v)}{\Delta u} = \frac{\partial}{\partial u} C(u, v).\end{aligned}$$

The procedure described above becomes

- Generate two independent uniform variates $u$ and $t$
- Set $v = {C_u}^{-1}(t)$, where ${C_u}^{-1}$ is any quasi-inverse of $C_u$
- The desired pair is $(u, v)$.

## Examples of Copulas

Several copula families are available that can incorporate the relationships between random variables. In what follows, the Gaussian copula and the Archimedean family of copulas are briefly discussed in a bivariate framework.

### *The Gaussian Copula*

The Gaussian or normal copula can be obtained by inversion from the well-known Gaussian bivariate distribution. It is defined as follows:

$$C^{Ga}(u_1, u_2; \rho) = \Phi(\phi^{-1}(u_1), \phi^{-1}(u_2); \rho),$$

where $\phi$ is the standard univariate Gaussian cumulative distribution and $\Phi(x, y; \rho)$ is the bivariate normal cumulative function with correlation $\rho$. Hence, the Gaussian or normal copula can be rewritten as follows:

$$C^{Ga}(u_1, u_2; \rho) = \int_{-\infty}^{\phi^{-1}(u_1)} \int_{-\infty}^{\phi^{-1}(u_2)} \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left\{\frac{-(s^2 - 2\rho st + t^2)}{2(1-\rho^2)}\right\} \mathrm{d}s\mathrm{d}t.$$

Variables with standard normal marginal distributions and this dependence structure are standard

bivariate normal variables with correlation coefficient $\rho$.

With $\rho = 0$ we obtain a very important special case of the Gaussian copula, which takes the form $C^{\perp}(u, v) = uv$ and is called *product copula*. The importance of this copula is related to the fact that two variables are independent if and only if their copula is the product copula.

### *The Archimedean Copulas*

The Archimedean copulas are characterized by their generator $\varphi$ through the following equation:

$$C(u_1, u_2) = \varphi^{-1}(\varphi(u_1) + \varphi(u_2))$$

for $u_1, u_2 \in [0, 1]$.

The following table presents three parametric Archimedean families that have gained interest in biostatistics, actuarial science, and management science, namely, the *Clayton copula,* the *Gumbel copula,* and the *Frank copula*.

| *Family* | $C(u_1, u_2)$ | *Generator* $\varphi(t)$ | *Comment* |
|---|---|---|---|
| Clayton | $\max\left[(u_1^{-\theta} + u_2^{-\theta} - 1)^{-1/\theta}, 0\right]$ | $\frac{1}{\theta}(t^{-\theta} - 1)$ | $\theta > 0$ |
| Gumbel | $exp\left[-\left((-ln(u_1))^{\theta} + (-ln(u_2))^{\theta}\right)^{1/\theta}\right]$ | $(-\ln(t))^{\theta}$ | $\theta \geq 1$ |
| Frank | $-\frac{1}{\theta} ln\left[1 + \frac{(e^{-\theta u_1} - 1)(e^{-\theta u_2} - 1)}{(e^{-\theta} - 1)}\right]$ | $-ln\left[\frac{(e^{-\theta t} - 1)}{(e^{-\theta} - 1)}\right]$ | $\theta > 0$ |

Each of the copulas presented in the preceding table is completely monotonic, and this allows for multivariate extension. If we assume that all pairs of random variables have the same $\phi$ and the same $\theta$ , the three copula functions can be extended by using the following relation:

$$C(u_{\downarrow}1, u_{\downarrow}2, \ldots, u_{\downarrow}n) = \varphi^{\uparrow}(-1)(\varphi(u_{\downarrow}1) + \varphi(u_{\downarrow}2) + \cdots + \varphi(u_{\downarrow}n))$$

for $u_1, u_2 \in [0, 1]$.

## Copulas and Dependence

Copulas are widely used in the study of dependence or association between random variables. Linear correlation (or Pearson's correlation) is the most frequently used measure of dependence in practice. Indeed, correlation is a natural measure of dependence in elliptical distributions (such as the multivariate normal and the multivariate $t$ distribution). However, most random variables are not jointly elliptically distributed, and using linear correlation as a measure of dependence in such situations might prove misleading. Two important measures of dependence, known as *Kendall's tau* and *Spearman's rho,* provide perhaps the best alternatives to the linear correlation coefficient as measures of dependence for nonelliptical distributions and can be expressed in terms of the underlying copula. Before presenting these measures and how they are related to copulas, we need to define the concordance concept.

### *Concordance*

Let $(x_1, y_1)$ and $(x_2, y_2)$ be two observations from a vector $(X, Y)$ of continuous random variables. Then $(x_1, y_1)$ and $(x_2, y_2)$ are said to be concordant if $(x_1 - x_2)(y_1 - y_2) > 0$ and discordant if $(x_1 - x_2)(y_1 - y_2) < 0$.

### *Kendall's tau*

Kendall's tau for a pair $(X, Y)$, distributed according to $H$, can be defined as the difference between the probabilities of concordance and discordance for two independent pairs $(X_1, Y_1)$ and $(X_2, Y_2)$, each with distribution $H$. This gives

$$\tau_{X,Y} = \Pr\{(X_1 - X_2)(Y_1 - Y_2) > 0\} - \Pr\{(X_1 - X_2)(Y_1 - Y_2) < 0\}.$$

The probabilities of concordance and discordance can be evaluated by integrating over the distribution of $(X_2, Y_2)$.

In terms of copulas, Kendall's tau becomes

$$\tau_{XY} = 4 \iint_{[0,1]^2} C(u, v)\, dC(u, v) - 1,$$

where $C$ is the copula associated to $(X, Y)$.

### *Spearman's rho*

Let $(X_1, Y_1)$, $(X_2, Y_2)$, and $(X_3, Y_3)$ be three independent random vectors with a common joint distribution function $H$ whose margins are $F$ and $G$. Consider the vectors $(X_1, Y_1)$ and $(X_2, Y_3)$—that is, a pair of vectors with the same margins, but one vector has distribution function $H$, while the components of the other are independent. The Spearman's rho coefficient associated to a pair

$(X, Y)$, distributed according to $H$, is defined to be proportional to the probability of concordance minus the probability of discordance for the two vectors $(X_1, Y_1)$ and $(X_2, Y_3)$:

$$\rho_{XY} = 3(\Pr\{(X_1 - X_2)(Y_1 - Y_3) > 0\} - \Pr\{(X_1 - X_2)(Y_1 - Y_3) < 0\}).$$

In terms of the copula $C$ associated to the pair $(X, Y)$, Spearman's rho becomes

$$\rho_{XY} = 12 \iint_{[0,1]^2} uv \, dC(u,v) - 3 = 12 \iint_{[0,1]^2} C(u,v) \, du dv - 3.$$

Note that if we let $U = F(X)$ and $V = G(Y)$, then

$$\rho_{XY} = 12 \iint_{[0,1]^2} uv \, dC(u,v) - 3 = 12E[UV] - 3 = \frac{E[UV] - 1/4}{1/12} = \frac{Cov(U,V)}{\sqrt{Var(U)}\sqrt{Var(V)}} = Corr(F(X),F(Y))$$

Kendall's tau and Spearman's rho are measures of dependence between two random variables. However, to extend to higher dimensions, we simply write pairwise measures in an $n \times n$ matrix in the same way as is done for linear correlation.

*Chiraz Labidi*

***See also*** Coefficients of Correlation, Alienation, and Determination; Correlation; Distribution

**Further Readings**

Dall'Aglio, G., Kotz, S. & Salinetti, G. (1991). *Advances in probability distributions with given margins*. Dordrecht, the Netherlands: Kluwer.

Embrechts, P., Lindskog F., & McNeil, A. (2001). *Modelling dependence with copulas and application to risk management*. Retrieved January 25, 2010, from http://www.risklab.ch/ftp/papers/DependenceWithCopulas.pdf

Frees, E. W., & Waldez, E. A. (1998). Understanding relationships using copulas. *North American Actuarial Journal, 2*(1).

Joe, H. (1997). *Multivariate models and dependence concepts*. London: Chapman & Hall.

Nelsen, R. B. (1999). *An introduction to copulas*. New York: Springer-Verlag.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges [N-dimensions' distribution functions and their margins]. *Publications de l'Institut Statistique de Paris, 8*, 229–231.

## Correction for Attenuation

Correction for attenuation (CA) is a method that allows researchers to estimate the relationship between two constructs as if they were measured perfectly reliably and free from random errors that occur in all observed measures. All research seeks to estimate the true relationship among constructs; because all measures of a construct contain random measurement error, the CA is especially important in order to estimate the relationships among constructs free from the effects of this error. It is called the CA because random measurement error attenuates, or makes smaller, the observed relationships between constructs. For correlations, the correction is as follows:

$$\rho_{xy} = \frac{r_{xy}}{\sqrt{r_{xx}}\sqrt{r_{yy}}}, \quad (1)$$

where $\delta_{xy}$ is the corrected correlation between variables $x$ and $y$, $r_{xy}$ is the observed correlation between variables $x$ and $y$, $r_{xx}$ is the reliability estimate for the $x$ variable, and $r_{yy}$ is the reliability estimate for the $y$ variable. For standardized mean differences, the CA is as follows:

$$\delta_{xy} = \frac{d_{xy}}{\sqrt{r_{yy}}}, \quad (2)$$

where $\delta_{xy}$ is the corrected standardized mean difference, $d_{xy}$ is the observed standardized mean difference, and $r_{yy}$ is the reliability for the continuous variable. In both equations, the observed effect size (correlation or standardized mean difference) is placed in the numerator of the right half of the equation, and the square root of the reliability estimate(s) is (are) placed in the denominator of the right half of the equation. The outcome from the left half of the equation is the estimate of the relationship between perfectly reliable constructs. For example, using Equation 1, suppose the observed

correlation between two variables is .25, the reliability for variable $X$ is .70, and the reliability for variable $Y$ is .80. The estimated true correlation between the two constructs is .25/(.70*.80) = .33. This entry describes the properties and typical uses of CA, shows how the CA equation is derived, and discusses advanced applications for CA.

## Properties

Careful examination of both Equations 1 and 2 reveals some properties of the CA. One of these is that the lower the reliability estimate, the higher the corrected effect size. Suppose that in two research contexts, $r_{xy} = .20$ both times. In the first context, $r_{xx} = r_{yy} = .50$, and in the second context, $r_{xx} = r_{yy} = .75$. The first context, with the lower reliability estimates, yields a higher corrected correlation ($\rho_{xy} = .40$) than the second research context ($\rho_{xy} = .27$) with the higher reliability estimates. An extension of this property shows that there are diminishing returns for increases in reliability; increasing the reliability of the two measures raises the corrected effect size by smaller and smaller amounts, as highly reliable measures begin to approximate the construct level or "true" relationship, where constructs have perfect reliability. Suppose now that $\rho_{xy} = .30$. If $r_{xx} = r_{yy}$, the corrected correlations when the reliabilities equal .50, .60, .70, .80, and .90, then the corrected correlations are .60, .50, .43, .38, and .33, respectively. Notice that while the reliabilities increase by uniform amounts, the corrected correlation is altered less and less.

Another property of Equation 1 is that it is not necessary to have reliability estimates of both variables $X$ and $Y$ in order to employ the CA. When correcting for only one variable, such as in applied research (or when reliability estimates for one variable are not available), a value of 1.0 is substituted for the reliability estimate of the uncorrected variable.

Another important consideration is the accuracy of the reliability estimate. If the reliability estimate is too high, the CA will underestimate the "true" relationship between the variables. Similarly, if the reliability estimate is too low, the CA will overestimate this relationship. Also, because correlations are bounded ($-1.0 \leq r_{xy} \leq 1.0$), it is possible to overcorrect outside these bounds if one (or both) of the reliability estimates is too low.

## Typical Uses

The CA has many applications in both basic and applied research. Because nearly every governing body in the social sciences advocates (if not requires) reporting effect sizes to indicate the strength of relationship, and because all effect sizes are attenuated because of measurement error, many experts recommend that researchers employ the CA when estimating the relationship between any pair of variables.

Basic research, at its essence, has a goal of estimating the true relationship between two constructs. The CA serves this end by estimating the relationship between measures that are free from measurement error, and all constructs are free from measurement error. In correlational research, the CA should be applied whenever two continuous variables are being related and when reliability estimates are available. However, when the reported effect size is either the $d$ value or the point-biserial correlation, the categorical variable (e.g., gender or racial group membership) is typically assumed to be measured without error; as such, no correction is made for this variable, and the CA is applied only to the continuous variable. In experimental research, the experimental manipulation is also categorical in nature and is typically assumed to be recorded without random measurement error. Again, the CA is only applied to the continuous dependent variable.

Although applied research also has a goal of seeking an estimate of the relationship between two variables at its core, often there are additional considerations. For example, in applied academic or employment selection research, the goal is not to estimate the relationship between the predictor and criterion constructs but to estimate the relationship between the predictor *measure* and the criterion construct. As such, the CA is not applied to the predictor. To do so would estimate the relationship between a predictor measure and a perfectly reliable criterion. Similarly, instances in which the relationship of interest is between a predictor *measure* and a criterion *measure* (e.g., between Medical College Admission Test scores and medical board examinations), the CA would not be appropriate, as the relationship of interest is between two observed measures.

## Derivation

In order to understand how and why the CA works, it is important to understand how the equation is derived. The derivation rests on two fundamental assumptions from classical test theory.

> Assumption 1: An individual's observed score on a particular measure of a trait is made up of two components: the individual's true standing on that trait and random measurement error.

This first assumption is particularly important to the CA. It establishes the equation that has become nearly synonymous with classical test theory:

$$X = T + e, \tag{3}$$

where $X$ is the observed score on a particular measure, $T$ is the true standing of the individual on the construct underlying the measure, and $e$ is random measurement error. Implicit in this equation is the fact that the construct in question is stable enough that there is a "true" score associated with it; if the construct being measured changes constantly, it is difficult to imagine a true score associated with that construct. As such, the CA is not appropriate when any of the constructs being measured are too unstable to be measured with a "true" score. Also revealed in this equation is that any departure from the true score in the observed score is entirely due to random measurement error.

> Assumption 2: Error scores are completely random and do not correlate with true scores or other error scores.

This assumption provides the basis for how scores combine in composites. Because observed test scores are composites of true and error scores (see Assumption 1), Assumption 2 provides guidance on how these components combine when one is calculating the correlation between two composite variables. Suppose there are observed scores for two variables, $A$ and $B$. Applying Equation 3,

$$X_a = T_a + e_a \tag{4}$$

$$X_b = T_b + e_b \tag{5}$$

where terms are defined as before, and subscripts denote which variable the terms came from. To calculate the correlation between two composite variables, it is easiest to use covariances. Assumption 2 defines how error terms in Equations 4 and 5 correlate with the other components

$$Cov(T_a, e_a) = Cov(T_b, e_b) = Cov(e_b, e_b) = 0, \tag{6}$$

where the $Cov()$ terms are the covariances among components in Equations 4 and 5. Suppose that one additional constraint is imposed: The observed score variance is 1.00. Applying Equations 3 and 6, this also means that the sum of true score variance and error score variance also equals 1.00. The covariance matrix between the components of variables $A$ and $B$ (from Equations 4, 5, and 6) is shown below:

$$\begin{bmatrix} Var(T_a) & 0 & Cov(T_a, T_b) & 0 \\ 0 & Var(e_a) & 0 & 0 \\ Cov(T_a, T_b) & 0 & Var(T_b) & 0 \\ 0 & 0 & 0 & Var(e_b) \end{bmatrix}, \tag{7}$$

where $Cov(T_a,T_b)$ is the covariance between true scores on variables $A$ and $B$, and the $Var()$ terms are the variances for the respective components defined earlier. The zero covariances between error scores and other variables are defined from Equation 6. It is known from statistical and psychometric theory that the correlation between two variables is equal to the covariance between the two variables divided by the square root of the variances of the two variables:

$$r_{ab} = \frac{Cov(A, B)}{\sqrt{Var_a}\sqrt{Var_b}}, \tag{8}$$

where $Var_a$ and $Var_b$ are variances for their respective variables. Borrowing notation defined in Equation 7 and substituting into equation 8, the correlation between observed scores on variables $A$ and $B$ is

$$r_{ab} = \frac{Cov(T_a, T_b)}{\sqrt{[Var(T_a) + Var(e_a)]}\sqrt{[Var(T_b) + Var(e_b)]}}. \tag{9}$$

Since the observed scores have a variance of 1.00 (because of the condition imposed earlier) and because of Equation 6, the correlation between true scores is equal to the covariance between true scores [i.e., $Cov(T_a,T_b) = r_{ab}$]. Making this substitution into Equation 9:

$$r_{ab} = \frac{r_{ab}}{\sqrt{[Var(T_a) + Var(e_b)]}\sqrt{[Var(T_b) + Var(e_b)]}}, \tag{10}$$

where all terms are as defined earlier. Note that the same term, $r_{ab}$, appears on both sides of Equation 10. By definition, $r_{ab} = r_{ab}$; mathematically, Equation 10 can be true only if the denominator is equal to 1.0. Because it was defined earlier that the variance of observed scores equals 1.0, Equation 9 can hold true. This requirement is relaxed for Equation 11 with an additional substitution.

Suppose now that a measure was free from measurement error, as it would be at the construct level. At the construct level, true relationships among variables are being estimated. As such, Greek letters are used to denote these relationships. If the true relationship between variables *A* and *B* is to be estimated, Equation 10 becomes

$$\rho_{ab} = \frac{r_{ab}}{\sqrt{[Var(T_a) + Var(e_a)]}\sqrt{[Var(T_b) + Var(e_b)]}}, \tag{11}$$

where $\rho_{ab}$ is the true correlation between variables *A* and *B*, as defined in Equation 1. Again, because variables are free from measurement error at the construct level (i.e., $Var(e_a) = Var(e_b) = 0$), Equation 11 becomes

$$\rho_{xy} = \frac{r_{xy}}{\sqrt{Var(T_x)}\sqrt{Var(T_y)}}. \tag{12}$$

Based on classical test theory, the reliability of a variable is defined to be the ratio of true score variance to observed score variance. In other words,

$$r_{xx} = \frac{Var(T)}{Var(X)}. \tag{13}$$

Because it was defined earlier that the observed score variance was equal to 1.0, Equation 13 can be rewritten to say that $r_{xx} = Var(T)$. Making this final substitution into Equation 12,

$$\rho_{ab} = \frac{r_{ab}}{\sqrt{r_{aa}}\sqrt{r_{bb}}}, \tag{14}$$

which is exactly equal to Equation 1 (save for the notation on variable names), or the CA. Though not provided here, a similar derivation can be used to obtain Equation 2, or the CA for *d* values.

## Advanced Applications

There are some additional applications of the CA under relaxed assumptions. One of these primary applications is when the correlation between error terms is not assumed to be zero. The CA under this condition is as follows:

$$\rho_{xy} = \frac{r_{xy} - r_{e_xe_y}\sqrt{1 - r_{xx}}\sqrt{1 - r_{yy}}}{\sqrt{rxx}\sqrt{ryy}}, \tag{15}$$

where $r_{e_xe_y}$ is the correlation between error scores for variables *X* and *Y*, and the other terms are as defined earlier.

It is also possible to correct part (or semipartial correlations) and partial correlations for the influences of measurement error. Using the standard formula for the partial correlation (in true score metric) between variables *X* and *Y*, controlling for *Z*, we get

$$\rho_{xy \cdot z} = \frac{\rho_{xy} - \rho_{xz}\rho_{yz}}{\sqrt{1 - (\rho_{xz})^2}\sqrt{1 - (\rho_{yz})^2)}}. \tag{16}$$

Substituting terms from Equation 1 into Equation 15 yields a formula to estimate the true partial correlation between variables *X* and *Y*, controlling for *Z*:

$$\rho_{xy \cdot z} = \frac{\frac{r_{xy}}{\sqrt{r_{xx}r_{yy}}} - \left(\frac{r_{xz}}{\sqrt{r_{xx}r_{zz}}}\right)\left(\frac{r_{yz}}{\sqrt{r_{yy}r_{zz}}}\right)}{\sqrt{1 - \left(\frac{r_{xz}}{\sqrt{r_{xx}r_{zz}}}\right)^2}\sqrt{1 - \left(\frac{r_{yz}}{\sqrt{r_{yy}r_{zz}}}\right)^2}} \tag{17}$$

Finally, it is also possible to compute the partial correlation between variables *X* and *Y*, controlling for *Z* while allowing the error terms to correlate:

$$\rho_{xyz} = \frac{\begin{array}{c} r_{zz}\left(r_{xy} - r_{e_x e_y}\sqrt{1 - r_{xx}}\sqrt{1 - r_{yy}}\right) - \\ \left(r_{xz} - r_{e_x e_z}\sqrt{1 - r_{xx}}\sqrt{1 - r_{zz}}\right) \\ \left(r_{yz} - r_{e_y e_z}\sqrt{1 - r_{yy}}\sqrt{1 - r_{zz}}\right) \end{array}}{\sqrt{r_{xx}r_{zz} - \left[r_{xz} - \left(r_{e_x e_z}\sqrt{1 - r_{xx}}\sqrt{1 - r_{zz}}\right)\right]^2}\sqrt{r_{yy}r_{zz} - \left[r_{yz} - \left(r_{e_y e_z}\sqrt{1 - r_{yy}}\sqrt{1 - r_{zz}}\right)\right]^2}} \quad (18)$$

Corrections for attenuation for part (or semipartial) correlations are also available under similar conditions as Equations 16 and 17.

*Matthew J. Borneman*

***See also*** Cohen's *d* Statistic; Correlation; Random Error; Reliability

**Further Readings**

Fan, X. (2003). Two approaches for correcting correlation attenuation caused by measurement error: Implications for research practice. *Educational & Psychological Measurement, 63,* 915–930.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences.* San Francisco: W. H. Freeman.

Schmidt, F. L., & Hunter, J. E. (1999). Theory testing and measurement error. *Intelligence, 27,* 183–198.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15,* 72–101.

Wetcher-Hendricks, D. (2006). Adjustments to the correction for attenuation. *Psychological Methods, 11,* 207–215.

Zimmerman, D. W., & Williams, R. H. (1977). The theory of test validity and correlated errors of measurement. *Journal of Mathematical Psychology, 16,* 135–152.

# Correlation

*Correlation* is a synonym for *association.* Within the framework of statistics, the term *correlation* refers to a group of indices that are employed to describe the magnitude and nature of a relationship between two or more variables. As a measure of correlation, which is commonly referred to as a *correlation coefficient,* is descriptive in nature, it cannot be employed to draw conclusions with regard to a cause–effect relationship between the variables in question. To infer cause and effect, it is necessary to conduct a controlled experiment involving an experimenter-manipulated independent variable in which subjects are randomly assigned to experimental conditions. Typically, data for which a correlation coefficient is computed are also evaluated with regression analysis. The latter is a methodology for deriving an equation that can be employed to estimate or predict a subject's score on one variable from the subject's score on another variable. This entry discusses the history of correlation and measures for assessing correlation.

## History

Although in actuality a number of other individuals had previously described the concept of correlation, Francis Galton, an English anthropologist, is generally credited with introducing the concept of correlation in a lecture on heredity he delivered in Great Britain in 1877. In 1896 Karl Pearson, an English statistician, further systematized Galton's ideas and introduced the now commonly employed product-moment correlation coefficient, which is represented by the symbol *r,* for interval-level or continuous data. In 1904 Charles Spearman, an English psychologist, published a method for computing a correlation for ranked data. During the 1900s numerous other individuals contributed to the theory and methodologies involved in correlation. Among the more notable contributors were William Gossett and Ronald Fisher, both of whom described the distribution of the *r* statistic; Udney Yule, who developed a correlational measure for categorical data, as well as working with Pearson to develop multiple correlation; Maurice Kendall, who developed alternative measures of correlation for ranked data; Harold Hotelling, who developed canonical correlation; and Sewall Wright, who developed path analysis.

## The Pearson Product-Moment Correlation

The Pearson product-moment correlation is the most commonly encountered bivariate measure of correlation. A bivariate correlation assesses the degree of relationship between two variables. The product-moment correlation describes the degree to which a linear relationship (the linearity is assumed) exists between one variable designated

as the *predictor variable* (represented symbolically by the letter $X$) and a second variable designated as the *criterion variable* (represented symbolically by the letter $Y$). The product-moment correlation is a measure of the degree to which the variables covary (i.e., vary in relation to one another). From a theoretical perspective, the product-moment correlation is the average of the products of the paired standard deviation scores of subjects on the two variables. The equation for computing the unbiased estimate of the population correlation is $r = (\sum z_x z_y)/(n-1)$.

The value $r$ computed for a sample correlation coefficient is employed as an estimate of $\rho$ (the lowercase Greek letter rho), which represents the correlation between the two variables in the underlying population. The value of $r$ will always fall within the range of $-1$ to $+1$ (i.e., $-1 \leq r \leq +1$). The absolute value of $r$ (i.e., $|r|$) indicates the strength of the linear relationship between the two variables, with the strength of the relationship increasing as the absolute value of $r$ approaches 1. When $r = \pm 1$, within the sample for which the correlation was computed, a subject's score on the criterion variable can be predicted perfectly from his or her score on the predictor variable. As the absolute value of $r$ deviates from 1 and moves toward 0, the strength of the relationship between the variables decreases, such that when $r = 0$, prediction of a subject's score on the criterion variable from his or her score on the predictor variable will not be any more accurate than a prediction that is based purely on chance.

The sign of $r$ indicates whether the linear relationship between the two variables is direct (i.e., an increase in one variable is associated with an increase in the other variable) or indirect (i.e., an increase in one variable is associated with a decrease on the other variable). The closer a positive value of $r$ is to $+1$, the stronger (i.e., more consistent) the direct relationship between the variables, and the closer a negative value of $r$ is to $-1$, the stronger the indirect relationship between the two variables. If the relationship between the variables is best described by a curvilinear function, it is quite possible that the value computed for $r$ will be close to zero. Because of the latter possibility, it is always recommended that a researcher construct a scatterplot of the data. A scatterplot is a graph that summarizes the two scores of each subject with a point in two-dimensional space. By examining the configuration of the scatterplot, a researcher can ascertain whether linear correlational analysis is best suited for evaluating the data.

Regression analysis is employed with the data to derive the equation of a *regression line* (also known as the *line of best fit*), which is the straight line that best describes the relationship between the two variables. To be more specific, a regression line is the straight line for which the sum of the squared vertical distances of all the points from the line is minimal. When $r = \pm 1$, all the points will fall on the regression line, and as the value of $r$ moves toward zero, the vertical distances of the points from the line increase.

The general equation for a regression line is $Y' = a + bX$, where $a$ = Y intercept, $b$ = the slope of the line (with a positive correlation yielding a positively sloped line, and a negative correlation yielding a negatively sloped line), $X$ represents a given subject's score on the predictor variable, and $Y'$ is the score on the criterion variable predicted for the subject.

An important part of regression analysis involves the analysis of residuals. A residual is the difference between the $Y'$ value predicted for a subject and the subject's actual score on the criterion variable. Use of the regression equation for predictive purposes assumes that subjects for whom scores are being predicted are derived from the same population as the sample for which the regression equation was computed. Although numerous hypotheses can be evaluated within the framework of the product-moment correlation and regression analysis, the most common null hypothesis evaluated is that the underlying population correlation between the variables equals zero. It is important to note that in the case of a large sample size, computation of a correlation close to zero may result in rejection of the latter null hypothesis. In such a case, it is critical that a researcher distinguish between statistical significance and practical significance, in that it is possible that a statistically significant result derived for a small correlation will be of no practical value; in other words, it will have minimal predictive utility.

A value computed for a product-moment correlation will be reliable only if certain assumptions regarding the underlying population distribution have not been violated. Among the assumptions

for the product-moment correlation are the following: (a) the distribution of the two variables is bivariate normal (i.e., each of the variables, as well as the linear combination of the variables, is distributed normally), (b) there is homoscedasticity (i.e., the strength of the relationship between the two variables is equal across the whole range of both variables), and (c) the residuals are independent.

## Alternative Correlation Coefficients

A common criterion for determining which correlation should be employed for measuring the degree of association between two or more variables is the levels of measurement represented by the predictor and criterion variables. The product-moment correlation is appropriate to employ when both variables represent either interval- or ratio-level data. A special case of the product-moment correlation is the *point-biserial correlation,* which is employed when one of the variables represents interval or ratio data and the other variable is represented on a dichotomous nominal scale (e.g., two categories, such as male and female). When the original scale of measurement for both variables is interval or ratio but scores on one of the variables have been transformed into a dichotomous nominal scale, the *biserial correlation* is the appropriate measure to compute. When the original scale of measurement for both variables is interval or ratio but scores on both of the variables have been transformed into a dichotomous nominal scale, the *tetrachoric correlation* is employed.

*Multiple correlation* involves a generalization of the product-moment correlation to evaluate the relationship between two or more predictor variables with a single criterion variable, with all the variables representing either interval or ratio data. Within the context of multiple correlation, partial and semipartial correlations can be computed. A *partial correlation* measures the relationship between two of the variables after any linear association one or more additional variables have with the two variables has been removed. A *semipartial correlation* measures the relationship between two of the variables after any linear association one or more additional variables have with one of the two variables has been removed. An extension of multiple correlation is *canonical correlation,* which involves assessing the relationship between a set of predictor variables (i.e., two or more) and a set of criterion variables.

A number of measures of association have been developed for evaluating data in which the scores of subjects have been rank ordered or the relationship between two or more variables is summarized in the format of a contingency table. Such measures may be employed when the data are presented in the latter formats or have been transformed from an interval or ratio format to one of the latter formats because one or more of the assumptions underlying the product-moment correlation are believed to have been saliently violated. Although, like the product-moment correlation, the range of values for some of the measures that will be noted is between $-1$ and $+1$, others may assume only a value between 0 and $+1$ or may be even more limited in range. Some alternative measures do not describe a linear relationship, and in some instances a statistic other than a correlation coefficient may be employed to express the degree of association between the variables.

Two methods of correlation that can be employed as measures of association when both variables are in the form of ordinal (i.e., rank order) data are *Spearman's rank order correlation* and *Kendall's tau. Kendall's coefficient of concordance* is a correlation that can be employed as a measure of association for evaluating three or more sets of ranks.

A number of measures of correlation or association are available for evaluating categorical data that are summarized in the format of a two-dimensional contingency table. The following measures can be computed when both the variables are dichotomous in nature: *phi coefficient* and *Yule's Q*. When both variables are dichotomous or one or both of the variables have more than two categories, the following measures can be employed: *contingency coefficient, Cramer's phi,* and *odds ratio.*

The *intraclass correlation* and *Cohen's kappa* are measures of association that can be employed for assessing interjudge reliability (i.e., degree of agreement among judges), the former being employed when judgments are expressed in the form of interval or ratio data, and the latter when the data are summarized in the format of a contingency table.

Measures of *effect size,* which are employed within the context of an experiment to measure the degree of variability on a dependent variable that can be accounted for by the independent variable, represent another type of correlational measure. Representative of such measures are *omega squared* and *eta-squared,* both of which can be employed as measures of association for data evaluated with an analysis of variance.

In addition to all the aforementioned measures, more complex correlational methodologies are available for describing specific types of curvilinear relationships between two or more variables. Other correlational procedures, such as *path analysis* and *structural equation modeling,* are employed within the context of *causal modeling* (which employs correlation data to evaluate hypothesized causal relationships among variables).

*David J. Sheskin*

***See also*** Bivariate Regression; Coefficients of Correlation, Alienation, and Determination; Least Squares, Methods of; Multiple Regression; Pearson Product-Moment Correlation Coefficient; Regression to the Mean; Residuals; Scatterplot

**Further Readings**

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Edwards, A. L. (1984). *An introduction to linear regression and correlation* (2nd ed.). New York: W. H. Freeman.

Howell, D. C. (2007). *Statistical methods for psychology* (6th ed.). Belmont, CA: Thomson/Wadsworth.

Montgomery, D. C., Peck, E. A., & Vining, G. (2006). *Introduction to linear regression analysis* (4th ed.). Hoboken, NJ: Wiley-Interscience.

Moore, D. S., & McCabe, G. P. (2009). *Introduction to the practice of statistics* (6th ed.). New York: W. H. Freeman.

Rosner, B. (2006). *Fundamentals of biostatistics* (6th ed.). Belmont, CA: Thomson-Brooks/Cole

Sheskin, D. J. (2007). *Handbook of parametric and nonparametric statistical procedures* (4th ed.). Boca Raton, FL: Chapman & Hall/CRC.

Zar, J. H. (1999). *Biostatistical analysis* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

# Correspondence Analysis

Correspondence analysis (CA) is a generalized principal component analysis tailored for the analysis of qualitative data. Originally, CA was created to analyze contingency tables, but CA is so versatile that it is used with a number of other data table types.

The goal of CA is to transform a data table into two sets of *factor scores:* one for the rows and one for the columns. The factor scores give the best representation of the similarity structure of the rows and the columns of the table. In addition, the factors scores can be plotted as maps, which display the essential information of the original table. In these maps, rows and columns are displayed as points whose coordinates are the factor scores and whose dimensions are called *factors*. It is interesting that the factor scores of the rows and the columns have the same variance, and therefore, both rows and columns can be conveniently represented in one single map.

The modern version of CA and its geometric interpretation comes from 1960s France and is associated with the French school of data analysis (*analyse des données*).

As a technique, it was often discovered (and rediscovered), and so variations of CA can be found under several different names, such as *dual-scaling, optimal scaling,* or *reciprocal averaging*. The multiple identities of CA are a consequence of its large number of properties: It can be defined as an optimal solution for many apparently different problems.

## Notations

Matrices are denoted with uppercase letters typeset in a boldface font; for example, $\mathbf{X}$ is a matrix. The elements of a matrix are denoted with a lowercase italic letter matching the matrix name, with indices indicating the row and column positions of the element; for example, $x_{i,j}$ is the element located at the $i$th row and $j$th column of matrix $\mathbf{X}$. Vectors are denoted with lowercase, boldface letters; for example, $\mathbf{c}$ is a vector. The elements of a vector are denoted with a lowercase italic letter matching the vector name and an index indicating the position

**Table 1** The Punctuation Marks of Six French Writers

| *Writer* | *Period* | *Comma* | *All Other Marks* |
|---|---|---|---|
| Rousseau | 7,836 | 13,112 | 6,026 |
| Chateaubriand | 53,655 | 102,383 | 42,413 |
| Hugo | 115,615 | 184,541 | 59,226 |
| Zola | 161,926 | 340,479 | 62,754 |
| Proust | 38,177 | 105,101 | 12,670 |
| Giraudoux | 46,371 | 58,367 | 14,299 |

*Source:* Adapted from Brunet, 1989.

of the element in the vector; for example $c_i$ is the $i$th element of $\mathbf{c}$. The italicized superscript $^T$ indicates that the matrix or vector is transposed.

## An Example: How Writers Punctuate

This example comes from E. Brunet, who analyzed the way punctuation marks were used by six French writers: Rousseau, Chateaubriand, Hugo, Zola, Proust, and Giraudoux. In the paper, Brunet gave a table indicating the number of times each of these writers used the period, the comma, and all the other marks (i.e., question mark, exclamation point, colon, and semicolon) grouped together. These data are reproduced in Table 1.

From these data we can build the original data matrix, which is denoted $\mathbf{X}$. It has $I = 6$ rows and $J = 3$ columns and is equal to

$$\mathbf{X} = \begin{bmatrix} 7836 & 13112 & 6026 \\ 53655 & 102383 & 42413 \\ 115615 & 184541 & 59226 \\ 161926 & 340479 & 62754 \\ 38177 & 105101 & 12670 \\ 46371 & 58367 & 14299 \end{bmatrix}. \quad (1)$$

In the matrix $\mathbf{X}$, the rows represent the authors and the columns represent types of punctuation marks. At the intersection of a row and a column, we find the number of a given punctuation mark (represented by the column) used by a given author (represented by the row).

## Analyzing the Rows

Suppose that the focus is on the authors and that we want to derive a map that reveals the similarities and differences in punctuation style among authors. In this map, the authors are points and the distances between authors reflect the proximity of style of the authors. So two authors close to each other punctuate in a similar way and two authors who are far from each other punctuate differently.

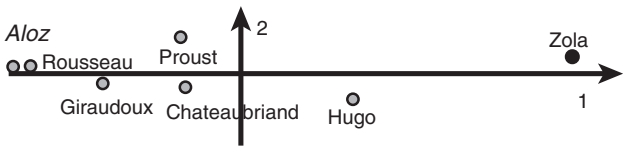



**Figure 1** Principal Components Analysis of Punctuation

*Notes:* Data are centered. Aloz, the fictitious alias for Zola, is a supplementary element. Even though Aloz punctuates the same way as Zola, Aloz is farther away from Zola than from any other author. The first dimension explains 98% of the variance. It reflects mainly the number of punctuation marks produced by the author.

### *A First Idea: Doing Principal Components Analysis*

A first idea is to perform a principal components analysis (PCA) on $\mathbf{X}$. The result is shown in Figure 1. The plot suggests that the data are quite unidimensional. And, in fact, the first component of this analysis explains 98% of the inertia (a value akin to variance). How to interpret this component? It seems related to the *number* of punctuation marks produced by each author. This interpretation is supported by creating a fictitious alias for Zola.

Suppose that, unbeknown to most historians of French literature, Zola wrote a small novel under the (rather transparent) pseudonym of Aloz. In this novel, he kept his usual way of punctuating, but because it was a short novel, he obviously produced a smaller number of punctuation marks than he did in his complete œuvre. Here is the (row) vector recording the number of occurrences of the punctuation marks for Aloz:

$$[2699 \quad 5675 \quad 1046]. \quad (2)$$

For ease of comparison, Zola's row vector is reproduced here:

$$[161926 \quad 340479 \quad 62754]. \quad (3)$$

So Aloz and Zola have the same punctuation style and differ only in their prolixity. A good analysis should reveal such a similarity of style, but as Figure 1 shows, PCA fails to reveal this similarity. In this figure, we have projected Aloz (as a supplementary element) in the analysis of the authors, and Aloz is, in fact, farther away from Zola than any other author. This example shows that using PCA to analyze the style of the authors is not a good idea because a PCA is sensitive mainly to the number of punctuation marks rather than to how punctuation is used. The "style" of the authors is, in fact, expressed by the *relative frequencies* of their use of the punctuation marks. This suggests that the data matrix should be transformed such that each author is described by the *proportion* of his usage of the punctuation marks rather than by the number of marks used. The transformed data matrix is called a *row profile* matrix. In order to obtain the row profiles, we divide each row by its sum. This matrix of row profiles is denoted **R**. It is computed as

$$\mathbf{R} = \operatorname{diag}\left\{\mathbf{X}\underset{J\times 1}{\mathbf{1}}\right\}^{-1}\mathbf{X} = \begin{bmatrix} .2905 & .4861 & .2234 \\ .2704 & .5159 & .2137 \\ .3217 & .5135 & .1648 \\ .2865 & .6024 & .1110 \\ .2448 & .6739 & .0812 \\ .3896 & .4903 & .1201 \end{bmatrix} \tag{4}$$

(where diag transforms a vector into a diagonal matrix with the elements of the vector on the diagonal, and $\frac{1}{J\times 1}$ is a $J \times 1$ vector of ones).

The "average writer" would be someone who uses each punctuation mark according to its proportion in the sample. The profile of this average writer would be the *barycenter* (also called *centroid, center of mass,* or *center of gravity*) of the matrix. Here, the barycenter of **R** is a vector with $J = 3$ elements. It is denoted **c** and computed as

$$\mathbf{c}^T = \underbrace{\left(\underset{1\times I}{\mathbf{1}}\times\mathbf{X}\times\underset{J\times 1}{\mathbf{1}}\right)^{-1}}_{\text{Inverse of the total of } \mathbf{X}} \times \underbrace{\underset{1\times I}{\mathbf{1}}\mathbf{X}}_{\text{Total of the columns of } \mathbf{X}} = [\,.2973 \quad .5642 \quad .1385\,]. \tag{5}$$

If all authors punctuate the same way, they all punctuate like the average writer. Therefore, in order to study the differences among authors, we need to analyze the matrix of *deviations* from the average writer. This matrix of deviations is denoted **Y,** and it is computed as

$$\mathbf{Y} = \mathbf{R} - \left(\underset{I\times 1}{\mathbf{1}}\times\mathbf{c}^T\right) = \begin{bmatrix} -.0068 & -.0781 & .0849 \\ -.0269 & -.0483 & .0752 \\ .0244 & -.0507 & .0263 \\ -.0107 & .0382 & -.0275 \\ -.0525 & .1097 & -.0573 \\ .0923 & -.0739 & -.0184 \end{bmatrix}. \tag{6}$$

### *Masses (Rows) and Weights (Columns)*

In CA, a mass is assigned to each row and a weight to each column. The mass of each row reflects its importance in the sample. In other words, the mass of each row is the proportion of this row in the total of the table. The masses of the rows are stored in a vector denoted **m,** which is computed as

$$\mathbf{m} = \underbrace{\left(\underset{1\times I}{\mathbf{1}}\times\mathbf{X}\times\underset{J\times 1}{\mathbf{1}}\right)^{-1}}_{\text{Inverse of the total of X}} \times \underbrace{\mathbf{X}\underset{J\times 1}{\mathbf{1}}}_{\text{Total of the rows of X}} = [.0189 \; .1393 \; .2522 \; .3966 \; .1094 \; .0835]^T. \tag{7}$$

From the vector **m,** we define the matrix of masses as $\mathbf{M} = \operatorname{diag}(\mathbf{m})$.

The weight of each column reflects its importance for *discriminating* among the authors. So the weight of a column reflects the information this column provides to the identification of a given row. Here, the idea is that columns that are used often do not provide much information, and column that are used rarely provide much information. A measure of how often a column is used is given by the proportion of times it is used, which is equal to the value of this column's component of the barycenter. Therefore the weight of a column is computed as the inverse of this column's component of the barycenter.

Specifically, if we denote by $\mathbf{w}$ the $J$ by 1 weight vector for the columns, we have

$$\mathbf{w} = [w_j] = \left[c_j^{-1}\right].$$

For our example, we obtain

$$\mathbf{w} = [w_j] = \left[c_j^{-1}\right] = \begin{bmatrix} \frac{1}{.2973} \\ \frac{1}{.5642} \\ \frac{1}{.1385} \end{bmatrix} = \begin{bmatrix} 3.3641 \\ 1.7724 \\ 7.2190 \end{bmatrix}. \quad (9)$$

From the vector $\mathbf{w}$, we define the matrix of weights as $\mathbf{W} = \text{diag}\{\mathbf{w}\}$.

*Generalized Singular Value Decomposition of Y*

Now that we have defined all these notations, CA boils down to a *generalized singular value decomposition* (GSVD) problem. Specifically, matrix $\mathbf{Y}$ is decomposed using the GSVD under the constraints imposed by the matrices $\mathbf{M}$ (masses for the rows) and $\mathbf{W}$ (weights for the columns):

$$\begin{aligned} \mathbf{Y} &= \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^T \quad \text{with:} \\ \mathbf{P}^T\mathbf{M}\mathbf{P} &= \mathbf{Q}^T\mathbf{W}\mathbf{Q} = \mathbf{I}, \end{aligned} \quad (10)$$

where $\mathbf{P}$ is the right singular vector, $\mathbf{Q}$ is the left singular vector, and $\boldsymbol{\Delta}$ is the diagonal matrix of the eigenvalues. From this we get

$$\begin{aligned} \mathbf{Y} = & \underbrace{\begin{bmatrix} 1.7962 & 0.9919 \\ 1.4198 & 1.4340 \\ 0.7739 & -0.3978 \\ -0.6878 & 0.0223 \\ -1.6801 & 0.8450 \\ 0.3561 & -2.6275 \end{bmatrix}}_{\mathbf{P}} \times \\ & \underbrace{\begin{bmatrix} .1335 & 0 \\ 0 & .0747 \end{bmatrix}}_{\boldsymbol{\Delta}} \times \\ & \underbrace{\begin{bmatrix} 0.1090 & -0.4114 & 0.3024 \\ -0.4439 & 0.2769 & 0.1670 \end{bmatrix}}_{\mathbf{Q}^T}. \end{aligned} \quad (11)$$

The rows of the matrix $\mathbf{X}$ are now represented by their factor scores (which are the projections of the observations onto the singular vectors). The row factor scores are stored in an $I = 3\ L = 2$ matrix (where $L$ stands for the number of nonzero singular values) denoted $\mathbf{F}$. This matrix is obtained as

$$\mathbf{F} = \mathbf{P}\boldsymbol{\Delta} = \begin{bmatrix} 0.2398 & 0.0741 \\ 0.1895 & 0.1071 \\ 0.1033 & -0.0297 \\ -0.0918 & 0.0017 \\ -0.2243 & 0.0631 \\ 0.0475 & -0.1963 \end{bmatrix}. \quad (12)$$

The variance of the factor scores for a given dimension is equal to the squared singular value of this dimension. (The variance of the observations is computed taking into account their masses.) Or equivalently, we say that the variance of the factor scores is equal to the eigenvalue of this dimension (i.e., the eigenvalue is the square of the singular value). This can be checked as follows:

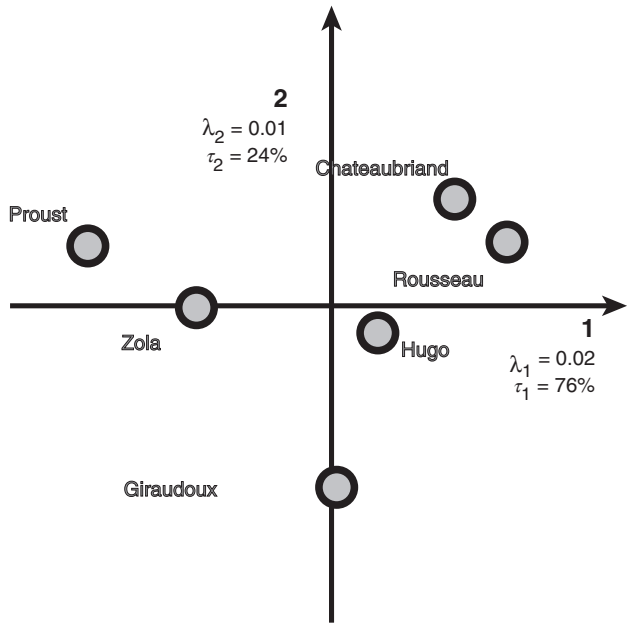



**Figure 2** Plot of the Correspondence Analysis of the Rows of Matrix **X**

*Notes:* The first two factors of the analysis for the rows are plotted (i.e., this is the matrix **F**). Each point represents an author. The variance of each factor score is equal to its eigenvalue.

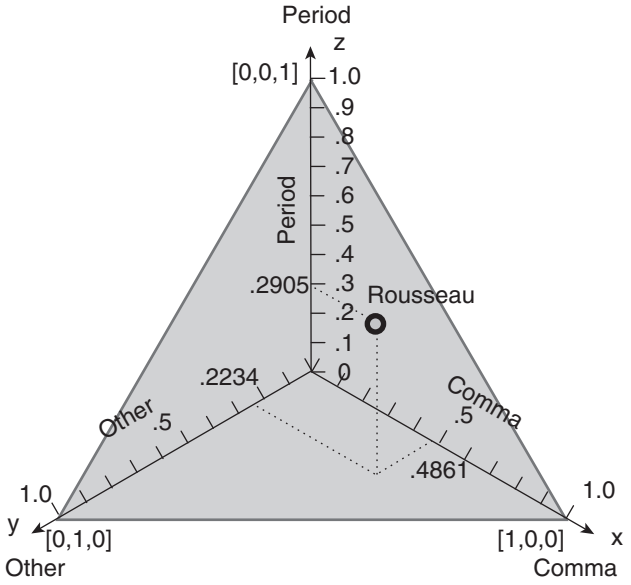



**Figure 3** In Three Dimensions, the Simplex Is a Two-Dimensional Triangle Whose Vertices Are the Vectors [100], [010], and [001]

$$\mathbf{F}^T\mathbf{M}\mathbf{F} = \mathbf{\Delta}^2 = \mathbf{L} = \begin{bmatrix} 0.1335^2 & 0 \\ 0 & 0.0747^2 \end{bmatrix} = \begin{bmatrix} 0.0178 & 0 \\ 0 & 0.0056 \end{bmatrix}. \quad (13)$$

We can display the results by plotting the factor scores as a map on which each point represents a row of the matrix $\mathbf{X}$ (i.e., each point represents an author). This is done in Figure 2. On this map, the first dimension seems to be related to time (the rightmost authors are earlier authors, and the leftmost authors are more recent), with the exception of Giraudoux, who is a very recent author. The second dimension singularizes Giraudoux. These factors will be easier to understand after we have analyzed the columns. This can be done by analyzing the matrix $\mathbf{X}^T$. Equivalently, it can be done by what is called *dual analysis*.

### *Geometry of the Generalized Singular Value Decomposition*

CA has a simple geometric interpretation. For example, when a row profile is interpreted as a vector, it can be represented as a point in a multidimensional space. Because the sum of a profile is equal to one, row profiles are, in fact points in a $J$ by 1 dimensional space.

Also, because the components of a row profile take value in the interval [0, 1], the points representing these row profiles can lie only in the subspace whose "extreme points" have one component equal to one and all other components equal to zero. This subspace is called a *simplex*. For example, Figure 3 shows the two-dimensional simplex

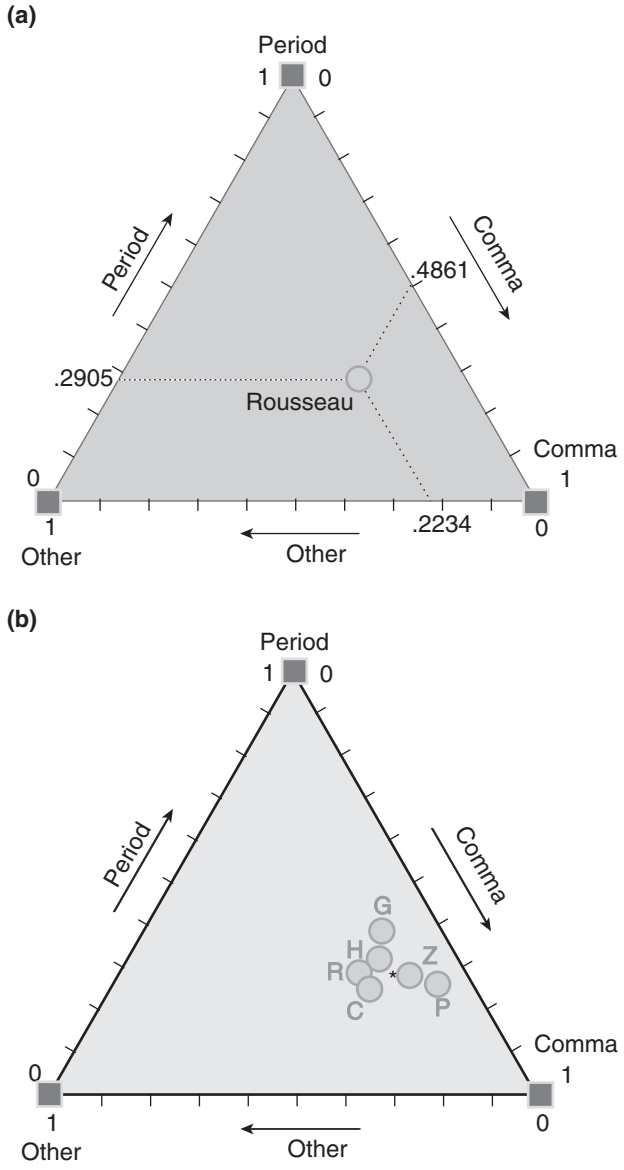



**Figure 4** The Simplex as a Triangle

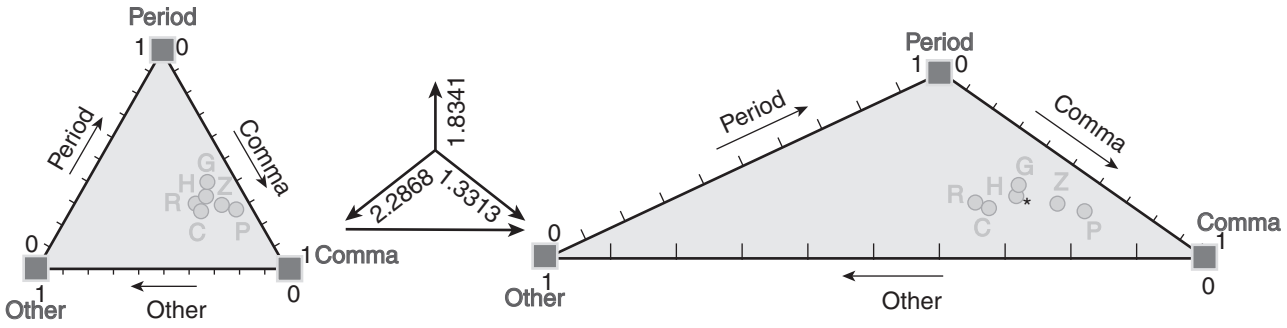



**Figure 5** Geometric Interpretation of the Column Weights

*Note:* Each side of the simplex is stretched by a factor equal to the square root of the weights.

corresponding to the subspace of all possible row profiles with three components. As an illustration, the point describing Rousseau (with coordinates equal to [.2905 .4861 .2234]) is also plotted. For this particular example, the simplex is an equilateral triangle and, so the three-dimensional row profiles can conveniently be represented as points on this triangle, as illustrated in Figure 4a, which shows the simplex of Figure 3 in two dimensions. Figure 4b shows all six authors and the barycenter.

The weights of the columns, which are used as constraints in the GSVD, also have a straightforward geometric interpretation. As illustrated in Figure 5, each side of the simplex is stretched by a quantity equal to the square root of the dimension it represents (we use the square root because we are interested in squared distances but not in squared weights, so using the square root of the weights ensures that the squared distances between authors will take into account the weight rather than the squared weights).

The masses of the rows are taken into account to find the dimensions. Specifically, the first factor is computed in order to obtain the maximum possible value of the sum of the masses times the squared projections of the authors' points (i.e., the projections have the largest possible variance). The second factor is constrained to be orthogonal (taking into account the masses) to the first one and to have the largest variance for the projections.

The remaining factors are computed with similar constraints. Figure 6 shows the stretched simplex, the author points, and the two factors (note that the origin of the factors is the barycenter of the authors).

The stretched simplex shows the whole space of the possible profiles. Figure 6 shows that the authors occupy a small portion of the whole space: They do not vary much in the way they punctuate. Also, the stretched simplex represents the columns as the vertices of the simplex: The columns are represented as row profiles with the column component being one and all the other components being zeros. This representation is called an *asymmetric* representation because the rows always have a dispersion smaller than (or equal to) the columns.

## Distance, Inertia, Chi-Square, and Correspondence Analysis

### *Chi-Square Distances*

In CA, the Euclidean distance in the stretched simplex is equivalent to a weighted distance in the

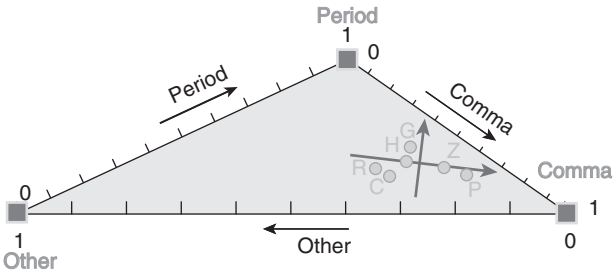



**Figure 6** Correspondence Analysis: The "Stretched Simplex" Along With the Factorial Axes

*Note:* The projections of the authors' points onto the factorial axes give the factor scores.

original space. For reasons that will be made more clear later, this distance is called the $\chi^2$ *distance*. The $\chi^2$ distance between two row profiles $i$ and $i'$ can be computed from the factor scores as

$$d^2_{i,i'} = \sum_{\ell}^{L} \left(f_{i,\ell} - f_{i',\ell}\right)^2 \tag{14}$$

or from the row profiles as

$$d^2_{i,i'} = \sum_{j}^{J} w_j \left(r_{i,j} - r_{i',j}\right)^2. \tag{15}$$

### *Inertia*

The variability of the row profiles relative to their barycenter is measured by a quantity—akin to variance—called *inertia* and denoted $\mathcal{I}$. The inertia of the rows to their barycenter is computed as the weighed sum of the squared distances of the rows to their barycenter. We denote by $d^2_{\mathrm{c},i}$ the (squared) distance of the $i$th row to the barycenter, computed as

$$d^2_{\mathrm{c},i} = \sum_{j}^{J} w_j (r_{i,j} - c_j)^2 = \sum_{l}^{L} f^2_{i,l} \tag{16}$$

where $L$ is the number of factors extracted by the CA of the table, [this number is smaller than or equal to min($I$, $J$) 1]. The inertia of the rows to their barycenter is then computed as

$$\mathcal{I} = \sum_{i}^{I} m_i d^2_{\mathrm{c},i}. \tag{17}$$

The inertia can also be expressed as the sum of the eigenvalues (see Equation 13):

$$\mathcal{I} = \sum_{\ell}^{L} \lambda_\ell \tag{18}$$

This shows that in CA, each factor extracts a portion of the inertia, with the first factor extracting the largest portion, the second factor extracting the largest portion left of the inertia, and so forth.

### *Inertia and the Chi-Square Test*

It is interesting that the inertia in CA is closely related to the chi-square test. This test is traditionally performed on a contingency table in order to test the independence of the rows and the columns of the table. Under independence, the frequency of each cell of the table should be proportional to the product of its row and column marginal probabilities. So if we denote by $x_{+,+}$ the grand total of matrix **X**, the expected frequency of the cell at the $i$th row and $j$th column is denoted $E_{i,j}$ and computed as

$$E_{i,j} = m_i c_j x_{+,+}. \tag{19}$$

The chi-square test statistic, denoted $\chi^2$, is computed as the sum of the squared difference between the actual values and the expected values, weighted by the expected values:

$$\chi^2 = \sum_{i,j} \frac{\left(x_{i,j} - E_{i,j}\right)^2}{E_{i,j}}. \tag{20}$$

When rows and columns are independent, $\chi^2$ follows a chi-square distribution with $(I - 1)(J - 1)$ degrees of freedom. Therefore, $\chi^2$ can be used to evaluate the likelihood of the row and columns independence hypothesis. The statistic $\chi^2$ can be rewritten to show its close relationship with the inertia of CA, namely:

$$\chi^2 = \mathcal{I} x_{+,+}. \tag{21}$$

This shows that CA analyzes—in orthogonal components—the pattern of deviations for independence.

## Dual Analysis

In a contingency table, the rows and the columns of the table play a similar role, and therefore the analysis that was performed on the rows can also be performed on the columns by exchanging the role of the rows and the columns. This is illustrated by the analysis of the columns of matrix **X**, or equivalently by the rows of the transposed matrix $\mathbf{X}^T$. The matrix of column profiles for $\mathbf{X}^T$ is called **O** (like cOlumn) and is computed as

$$\mathbf{O} = \operatorname{diag}\left\{\mathbf{X}^T \underset{I \times 1}{\mathbf{1}}\right\}^{-1} \mathbf{X}^T. \qquad (22)$$

The matrix of the deviations to the barycenter is called $\mathbf{Z}$ and is computed as

$$\mathbf{Z} = \mathbf{O} - \left(\underset{I \times 1}{\mathbf{1}} \times \mathbf{m}^T\right) = \begin{bmatrix} -.0004 & -.0126 & .0207 & -.0143 & -.0193 & .0259 \\ -.0026 & -.0119 & -.0227 & .0269 & .0213 & -.0109 \\ .0116 & .0756 & .0478 & -.0787 & -.0453 & -.0111 \end{bmatrix}.$$

Weights and masses of the columns analysis are the inverse of their equivalent for the row analysis. This implies that the punctuation marks factor scores are obtained from the GSVD with the constraints imposed by the two matrices $\mathbf{W}^{-1}$ (masses for the rows) and $\mathbf{M}^{-1}$ (weights for the columns; compare with Equation 10):

$$\mathbf{Z} = \mathbf{U}\mathbf{D}\mathbf{V}^T \qquad \text{with:} \qquad \mathbf{U}^T\left(\mathbf{W}^{-1}\right)\mathbf{U} = \mathbf{V}^T\left(\mathbf{M}^{-1}\right)\mathbf{V} = \mathbf{I}\,. \qquad (23)$$

This gives

$$\mathbf{Z} = \underbrace{\begin{bmatrix} 0.3666 & -1.4932 \\ -0.7291 & 0.4907 \\ 2.1830 & 1.2056 \end{bmatrix}}_{\mathbf{U}} \times \underbrace{\begin{bmatrix} .1335 & 0 \\ 0 & .0747 \end{bmatrix}}_{\boldsymbol{\Delta}} \times \underbrace{\begin{bmatrix} 0.0340 & 0.1977 & 0.1952 & -0.2728 & -0.1839 & 0.0298 \\ 0.0188 & 0.1997 & -0.1003 & 0.0089 & 0.0925 & -0.2195 \end{bmatrix}}_{\mathbf{V}^T}. \qquad (24)$$

The factor scores for the punctuation marks are stored in a $J = 3 \times L = 2$ matrix called $\mathbf{G}$, which is computed in the same way $\mathbf{F}$ was computed (see Equation 12). So $\mathbf{G}$ is computed as

$$\mathbf{G} = \mathbf{U}\boldsymbol{\Delta} = \begin{bmatrix} 0.0489 & -0.1115 \\ -0.0973 & 0.0367 \\ 0.2914 & 0.0901 \end{bmatrix}. \qquad (25)$$

*Transition Formula*

A comparison of Equation 24 with Equation 11 shows that the singular values are the same for both analyses. This means that the inertia (i.e., the square of the singular value) extracted by each factor is the same for both analyses. Because the variance extracted by the factors can be added to obtain the total inertia of the data table, this also means that each analysis is decomposing the same inertia, which here is equal to

$$\mathcal{I} = .1335^2 + .747^2 = .0178 + .0056 = 0.0234\,. \qquad (26)$$

Also, the generalized singular decomposition of one set (say, the columns) can be obtained from the other one (say, the rows). For example, the generalized singular vectors of the analysis of the columns can be computed directly from the analysis from the rows as

$$\mathbf{U} = \mathbf{W}\mathbf{Q}. \qquad (27)$$

Combining Equations 27 and 25 shows that the factors for the rows of $\mathbf{Z}$ (i.e., the punctuation marks) can be obtained directly from the singular value decomposition of the authors matrix (i.e., the matrix $\mathbf{Y}$) as

$$\mathbf{G} = \mathbf{W}\mathbf{Q}\boldsymbol{\Delta}. \qquad (28)$$

As a consequence, we can, in fact, find directly the

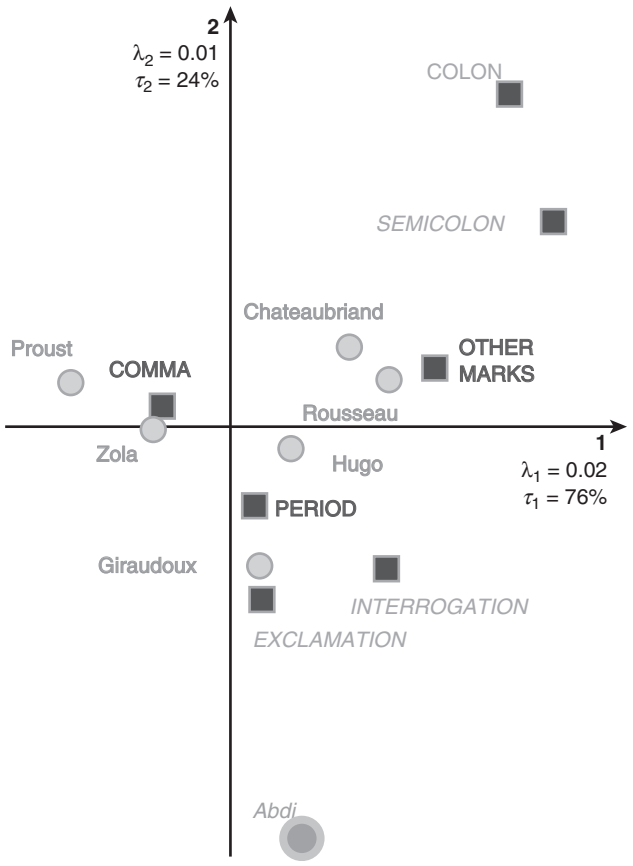



**Figure 7** Correspondence Analysis of the Punctuation of Six Authors

*Notes:* Comma, period, and other marks are active columns; Rousseau, Chateaubriand, Hugo, Zola, Proust, and Giraudoux are active rows. Colon, semicolon, interrogation, and exclamation are supplementary columns; Abdi is a supplementary row.

factor scores of the columns from their profile matrix (i.e., the matrix **O**), and from the factor scores of the rows. Specifically, the equation that gives the values of **O** from **F** is

$$\mathbf{G} = \mathbf{O}\mathbf{F}\boldsymbol{\Delta}^{-1}, \tag{29}$$

and conversely, **F** could be obtained from **G** as

$$\mathbf{F} = \mathbf{R}\mathbf{G}\boldsymbol{\Delta}^{-1}. \tag{30}$$

These equations are called *transition formulas* from the rows to the columns (and vice versa) or simply the transition formulas.

### *One Single Generalized Singular Value Decomposition for Correspondence Analysis*

Because the factor scores obtained for the rows and the columns have the same variance (i.e., they have the same "scale"), it is possible to plot them in the same space. This is illustrated in Figure 7. The symmetry of the rows and the columns in CA is revealed by the possibility of directly obtaining the factors scores from one single GSVD. Specifically, let $\mathbf{D}_m$ and $\mathbf{D}_c$ denote the diagonal matrices with the elements of **m** and **c**, respectively, on the diagonal, and let **N** denote the matrix **X** divided by the sum of all its elements. This matrix is called a *stochastic* matrix: All its elements are larger than zero, and their sum is equal to one. The factor scores for the rows and the columns are obtained from the following GSVD:

$$\begin{aligned} \left(\mathbf{N} - \mathbf{m}\mathbf{c}^T\right) &= \mathbf{S}\boldsymbol{\Delta}\mathbf{T}^T \quad \text{with} \\ \mathbf{S}^T\mathbf{D}_{\mathbf{m}}^{-1}\mathbf{S} &= \mathbf{T}^T\mathbf{D}_{\mathbf{c}}^{-1}\mathbf{T} = \mathbf{I}\,. \end{aligned} \tag{31}$$

The factor scores for the rows (**F**) and the columns (**G**) are obtained as

$$\mathbf{F} = \mathbf{D}_{\mathbf{m}}^{-1}\mathbf{S}\boldsymbol{\Delta} \quad \text{and} \quad \mathbf{G} = \mathbf{D}_{\mathbf{c}}^{-1}\mathbf{T}\boldsymbol{\Delta}. \tag{32}$$

## Supplementary Elements

Often in CA we want to know the position in the analysis of rows or columns that were not analyzed. These rows or columns are called illustrative or supplementary rows or columns (or supplementary observations or variables). By contrast with the appellation of supplementary (i.e., not used to compute the factors), the *active* elements are those used to compute the factors. Table 2 shows the punctuation data with four additional columns giving the detail of the "other punctuation marks" (i.e., the exclamation point, the question mark, the semicolon, and the colon). These punctuation marks were not analyzed for two reasons: First, these marks are used rarely, and therefore they would distort the factor space, and second, the "other" marks comprises all the other marks, and therefore to analyze them with "other" would be redundant. There is also a new author in Table 2: We counted the marks used by a different author, namely, Hervé Abdi in the first chapter of his 1994

**Table 2** Number of Punctuation Marks Used by Six Major French Authors

| | *Active Elements* | | | | | *Supplementary Elements* | | | |
|---|---|---|---|---|---|---|---|---|---|
| | *Period* | *Comma* | *Other Marks* | $x_{i,+}$ | $m = \frac{x_{i+}}{x_{++}}$ | *Exclamation* | *Question* | *Semicolon* | *Colon* |
| Rousseau | 7,836 | 13,112 | 6,026 | **26,974** | **.0189** | 413 | 1,240 | 3,401 | 972 |
| Chateaubriand | 53,655 | 102,383 | 42,413 | **198,451** | **.1393** | 4,669 | 4,595 | 19,354 | 13,795 |
| Hugo | 115,615 | 184,541 | 59,226 | **359,382** | **.2522** | 19,513 | 9,876 | 22,585 | 7,252 |
| Zola | 161,926 | 340,479 | 62,754 | **565,159** | **.3966** | 24,025 | 10,665 | 18,391 | 9,673 |
| Proust | 38,117 | 105,101 | 12,670 | **155,948** | **.1094** | 2,756 | 2,448 | 3,850 | 3,616 |
| Giraudoux | 46,371 | 58,367 | 14,229 | **119,037** | **.0835** | 5,893 | 5,042 | 1,946 | 1,418 |
| $x_{+j}$ | **423,580** | **803,983** | **197,388** | **1,424,951** | | | | | |
| $\mathbf{W}^T = \frac{x_{i+}}{x_{++}}$ | **3.3641** | **1.7724** | **7.2190** | $x_{++}$ | | | | | |
| $\mathbf{c}^T = \frac{x_{i+}}{x_{++}}$ | **.2973** | **5642** | **.1385** | | | | | | |
| Abdi (Chapter 1) | 216 | 139 | 26 | | | | | | |

*Source:* Adapted from Brunet, 1989.

*Notes:* The exclamation point, question mark, semicolon, and colon are supplementary columns. Abdi (1994) Chapter 1 is a supplementary row. Notations: $x_{i+}$ = sum of the $i$th row; $x + j$ = sum of the $j$th column; $x_{++}$ = grand total.

book called *Les réseaux de reurores*. This author was not analyzed because the data are available for only one chapter (not his complete work) and also because this author is not a literary author.

The values of the projections on the factors for the supplementary elements are computed from the transition formula. Specifically, a supplementary row is projected into the space defined using the transition formula for the active rows (cf. Equation 30) and replacing the active row profiles by the supplementary row profiles. So if we denote by $\mathbf{R}_{\text{sup}}$ the matrix of the supplementary row profiles, then $\mathbf{F}_{\text{sup}}$—the matrix of the supplementary row factor scores—is computed as

$$\mathbf{F}_{\text{sup}} = \mathbf{R}_{\text{sup}} \times \mathbf{G} \times \mathbf{\Delta}^{-1}. \quad (33)$$

Table 3 provides factor scores and descriptives for the rows.

For example, the factor scores of the author Abdi are computed as

$$\mathbf{F}_{\text{sup}} = \mathbf{R}_{\text{sup}}\mathbf{G}\mathbf{\Delta}^{-1} = [0.0908 - 0.5852]. \quad (34)$$

Supplementary columns are projected into the factor space using the transition formula from the active rows (cf. Equation 29) and replacing the active column profiles by the supplementary column profiles. If we denote by $\mathbf{O}_{\text{sup}}$ the supplementary column profile matrix, then $\mathbf{G}_{\text{sup}}$, the matrix of the supplementary column factor scores, is computed as

$$\mathbf{G}_{\text{sup}} = \mathbf{O}_{\text{sup}}\mathbf{F}\mathbf{\Delta}^{-1}. \quad (35)$$

Table 4 gives the factor scores for the supplementary elements.

## Little Helpers: Contributions and Cosines

Contributions and cosines are coefficients whose goal is to facilitate the interpretation. The contributions identify the important elements for a given factor, whereas the (squared) cosines identify the factors important for a given element. These coefficients express importance as the proportion of something in a total. The contribution is the ratio of the weighted squared projection of an element on a factor to the sum of the weighted projections of all the elements for this factor (which happens to be the eigenvalue of this factor). The squared cosine is the ratio of the squared projection of an element on a factor to the sum of the projections of this element on all the factors (which happens to be the squared distance from this point to the barycenter). Contributions and squared cosines are proportions that vary between 0 and 1.

**Table 3** Factor Scores, Contributions, and Cosines for the Rows

| *Axis* | λ | % | *Rousseau* | *Chateaubriand* | *Hugo* | *Zola* | *Proust* | *Giraudoux* | *Abdi (Chapter 1)* |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | *Factor Scores* | | | | |
| 1 | .0178 | 76 | 0.2398 | 0.1895 | 0.1033 | –0.0918 | –0.2243 | 0.0475 | –0.0908 |
| 2 | .0056 | 24 | 0.0741 | 0.1071 | –0.0297 | 0.0017 | 0.0631 | –0.1963 | 0.5852 |
| | | | | | *Contributions* | | | | |
| 1 | | | 0.0611 | 0.2807 | 0.1511 | *0.1876* | *0.3089* | 0.0106 | – |
| 2 | | | 0.0186 | 0.2864 | *0.0399* | 0.0002 | 0.0781 | *0.5767* | – |
| | | | | | *Cosines* | | | | |
| 1 | | | 0.9128 | 0.7579 | 0.9236 | 0.9997 | 0.9266 | 0.0554 | 0.0235 |
| 2 | | | 0.0872 | 0.2421 | 0.0764 | 0.0003 | 0.0734 | 0.9446 | 0.9765 |
| | | | | | *Squared distances to grand barycenter* | | | | |
| – | – | – | 0.0630 | 0.0474 | 0.0116 | 0.0084 | 0.0543 | 0.0408 | 0.3508 |

*Notes:* Negative contributions are shown in italic. Abdi (1994) Chapter 1 is a supplementary row.

**Table 4** Factor Scores, Contributions, and Cosines for the Columns

| *Axis* | λ | % | *Period* | *Comma* | *Other Marks* | *Exclamation* | *Question* | *Semicolon* | *Colon* |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | *Factor scores* | | | | |
| 1 | .0178 | 76 | –0.0489 | 0.0973 | –0.2914 | –0.0596 | –0.1991 | –0.4695 | –0.4008 |
| 2 | .0056 | 24 | 0.1115 | –0.0367 | -0.0901 | 0.2318 | 0.2082 | –0.2976 | –0.4740 |
| | | | | | *Contributions* | | | | |
| 1 | | | 0.0399 | 0.2999 | 0.6601 | – | – | – | – |
| 2 | | | 0.6628 | *0.1359* | *0.2014* | – | – | – | – |
| | | | | | *Cosines* | | | | |
| 1 | | | 0.1614 | 0.8758 | 0.9128 | 0.0621 | 0.4776 | 0.7133 | 0.4170 |
| 2 | | | 0.8386 | 0.1242 | 0.0872 | 0.9379 | 0.5224 | 0.2867 | 0.5830 |
| | | | | | *Squared distances to grand barycenter* | | | | |
| – | – | | 0.0148 | 0.0108 | 0.0930 | 0.0573 | 0.0830 | 0.3090 | 0.3853 |

*Notes:* Negative contributions are shown in italic. Exclamation point, question mark, semicolon, and colon are supplementary columns.

The squared *cosines*, denoted $h$, between row $i$ and factor $l$ (and between column $j$ and factor $l$) are obtained as

$$h_{i,\ell} = \frac{f_{i,\ell}^2}{\sum_{\ell} f_{i,\ell}^2} = \frac{f_{i,\ell}^2}{d_{\mathbf{c},i}^2} \quad \text{and} \quad h_{j,\ell} = \frac{g_{j,\ell}^2}{\sum_{\ell} f_{j,\ell}^2} = \frac{g_{j,\ell}^2}{d_{\mathbf{r},j}^2}. \quad (36)$$

Squared cosines help in locating the factors important for a given observation. The contributions, denoted $b$, of row $i$ to factor $l$ and of column $j$ to factor $l$ are obtained as

$$b_{i,\ell} = \frac{m_i f_{i,\ell}^2}{\sum_{i} m_i f_{i,\ell}^2} = \frac{m_i f_{i,\ell}^2}{\lambda_\ell} \quad \text{and} \quad b_{j,\ell} = \frac{c_j g_{j,\ell}^2}{\sum_{j} c_j f_{j,\ell}^2} = \frac{c_i g_{j,\ell}^2}{\lambda_\ell}. \quad (37)$$

Contributions help locating the observations important for a given factor. An often used rule of thumb is to consider that the important contributions are larger than the average contribution, which is equal to the number of elements (i.e., $\frac{1}{I}$ for the rows and $\frac{1}{J}$ for the columns). A dimension is then interpreted by opposing the

positive elements with large contributions to the negative elements with large contributions. Cosines and contributions for the punctuation example are given in Tables 3 and 4.

### *Multiple Correspondence Analysis*

CA works with a contingency table that is equivalent to the analysis of two nominal variables (i.e., one for the rows and one for the columns). Multiple CA (MCA) is an extension of CA that analyzes the pattern of relationships among several nominal variables. MCA is used to analyze a set of observations described by a set of nominal variables. Each nominal variable comprises several levels, and each of these levels is coded as a binary variable. For example, gender (F vs. M) is a nominal variable with two levels. The pattern for a male respondent will be [0 1], and for a female respondent, [1 0]. The complete data table is composed of binary columns with one and only one column, per nominal variable, taking the value of 1.

MCA can also accommodate quantitative variables by recoding them as "bins." For example, a score with a range of $-5$ to $+5$ could be recoded as a nominal variable with three levels: less than 0, equal to 0, or more than 0. With this schema, a value of 3 will be expressed by the pattern 0 0 1. The coding schema of MCA implies that each row has the same total, which for CA implies that each row has the same *mass*.

Essentially, MCA is computed by using a CA program on the data table. It can be shown that the binary coding scheme used in MCA creates artificial factors and therefore artificially reduces the inertia explained. A solution for this problem is to correct the eigenvalues obtained from the CA program.

*Hervé Abdi and Lynne J. Williams*

***See also*** Barycentric Discriminant Analysis; Canonical Correlation Analysis; Categorical Variable; Chi-Square Test; Coefficient Alpha; Data Mining; Descriptive Discriminant Analysis; Discriminant Analysis; Exploratory Data Analysis; Exploratory Factor Analysis; Guttman Scaling; Matrix Algebra; Principal Components Analysis; R

**Further Readings**

Benzécri, J. P. (1973). *L'analyse des données* [Data analysis] (2 vol.). Paris: Dunod.

Brunet, E. (1989). Faut-il ponderer les données linguistiques [Weighting linguistics data]. *CUMFID, 16,* 39–50.

Clausen, S. E. (1998). *Applied correspondence analysis.* Thousand Oaks, CA: Sage.

Escofier, B., & Pages, J. (1998). *Analyses factorielles simples et multiples* [Simple and multiple factor analysis]. Paris: Dunod.

Greenacre, M. J. (1984). *Theory and applications of correspondence analysis.* London: Academic Press.

Greenacre, M. J. (2007). *Correspondence analysis in practice* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Greenacre, M. J., & Blasius, J. (Eds.). (2007). *Multiple correspondence analysis and related methods.* Boca Raton, FL: Chapman & Hall/CRC.

Hwang, H., Tomiuk, M. A., & Takane, Y. (2009). Correspondence analysis, multiple correspondence analysis and recent developments. In R. Millsap & A. Maydeu-Olivares (Eds.), *Handbook of quantitative methods in psychology* (pp. 243–263). London: Sage.

Lebart, L., & Fenelon, J. P. (1971). *Statistiques et informatique appliquees* [Applied statistics and computer science]. Paris: Dunod.

Weller, S. C., & Romney, A. K. (1990). *Metric scaling: Correspondence analysis.* Thousand Oaks, CA: Sage.

## Correspondence Principle

The correspondence principle is generally known as the Bohr correspondence principle (CP), for Niels Bohr. It is considered one of Bohr's greatest contributions to physics, along with his derivation of the Balmer formula. Bohr's leading idea is that classical physics, though limited in scope, is indispensable for the understanding of quantum physics. The idea that old science is "indispensable" to the understanding of new science is in fact the main theme in using the concept of correspondence; therefore, the CP can be defined as the principle by which new theories of science (physics in particular) can relate to previously accepted theories in the field by means of approximation at a certain limit. Historically, Max Planck had introduced the concept in 1906. Bohr's first handling of the concept was in his first paper after World War I, in which he showed that quantum

formalism would lead to classical physics when $n \to \infty$, where n is the quantum number. Although there were many previous uses of the concept, the important issue here is not to whom the concept can be attributed, but an understanding of the various ways that it can be used in scientific and philosophic research.

The principle is important for the continuity in science. There are two ways of thinking about such continuity. A theory $T$ covers a set of observations $S$. A new observation $s_1$ is detected. $T$ cannot explain $s_1$. Scientists first try to adapt $T$ to be able to account for $s_1$. But if $T$ is not in principle able to explain $s_1$, then scientists will start to look for another theory, $T^*$, that can explain $S$ and $s_1$. The scientist will try to derive $T^*$ by using CP as a determining factor. In such a case, $T^*$ should lead to $T$ at a certain limit.

Nonetheless, sometimes there may be a set of new observations, $S_1$, for which it turns out that a direct derivation of $T^*$ from $T$ that might in principle account for $S_1$ is not possible or at least does not seem to be possible. Then the scientist will try to suggest $T^*$ separately from the accepted set of boundary conditions and the observed set of $S$ and $S_1$. But because $T$ was able to explain the set of it is highly probable that $T$ has a certain limit of correct assumptions that led to its ability to explain $S$. Therefore, any new theory $T^*$ that would account for $S$ and $S_1$ should resemble $T$ at a certain limit. This can be obtained by specifying a certain correspondence limit at which the new formalism of $T^*$ will lead to the old formalism of $T$.

These two ways of obtaining $T^*$ are the general forms of applying the correspondence principle. Nevertheless, the practice of science presents us with many ways of connecting $T^*$ to $T$ or parts of it. Hence it is important to discuss the physicists' different treatments of the CP. Moreover, the interpretation of CP and the implications of using CP will determine our picture of science and the future development of science; hence, it is important to discuss the philosophical implications of CP and the different philosophical understandings of the concept.

## Formal Correspondence

In the current state of the relation between modern physics and classical physics, there are four kinds of formal correspondence between modern and classical physics.

### *Old Correspondence Principle (Numerical Correspondence)*

Planck stressed the relation between his "radical" assumption of discrete energy levels that are proportional to frequency, and the classical theory. He insisted that the terms in the new equation refer to the very same classical properties. He formulated the CP so that the numerical value of

$$\lim_{h \to 0}[Quantumphysics] = [Classicalphysics]$$

He demonstrated that the radiation law for the energy density at frequency $\nu$,

$$u(\upsilon) = \frac{8\pi h\upsilon^3}{c^3(e^{h\upsilon/kT} - 1),} \tag{1}$$

corresponds numerically in the limit $h \to 0$ to the classical Rayleigh–Jeans law:

$$u(\upsilon) = \frac{8\pi kT\upsilon^2}{c^3}, \tag{2}$$

where $k$ is Boltzmann's constant, $T$ is the temperature, and $c$ is the speed of light. This kind of correspondence entails that the new theory should resemble the old one not just at the mathematical level but also at the conceptual level.

### *Configuration Correspondence Principle (Law Correspondence)*

The configuration correspondence principle claims that the laws of new theories should correspond to the laws of the old theory. In the case of quantum and classical physics, quantum laws correspond to the classical laws when the probability density of the quantum state coincides with the classical probability density. Take, for example, a harmonic oscillator that has a classical probability density

$$P_C(x) = {}^{1}/_{(\pi\sqrt{x_0^2 - x^2})}, \tag{3}$$

where $x$ is the displacement. Now if we superimpose the plot of this probability onto that of the quantum probability density $|\psi_n|^2$ of the eigenstates of the system and take (the quantum number) $n \to \infty$, we will obtain Figure 1 below. As

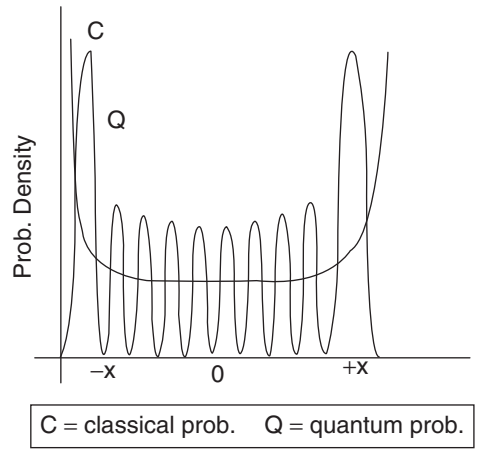



**Figure 1** Classical Versus Quantum Probability Density

Richard Liboff, a leading expert in the field, has noted, the classical probability density $P_C$ does not follow the quantum probability density $|\psi_n|^2$. Instead, it follows the local average in the limit of large quantum numbers *n:*

$$P_C(x) = \langle P_Q(x) \rangle = \left\langle |\psi_n|^2 \right\rangle = \tfrac{1}{2\varepsilon} \int_{x-\varepsilon}^{x+\varepsilon} |\psi_n(y)|^2 dy. \quad (4)$$

### *Frequency Correspondence Principle*

The third type of correspondence is the officially accepted form of correspondence that is known in quantum mechanics books as the Bohr Correspondence Principle. This claims that the classical results should emerge as a limiting case of the quantum results in the limits $n \to \infty$ (the quantum number) and $h \to 0$ (Planck's constant). Then in the case of frequency, the quantum value should be equal to the classical value, i.e., $\upsilon_Q = \upsilon_C$. In most cases in quantum mechanics, the quantum frequency coalesces with the classical frequency in the limit $n \to \infty$ and $h \to 0$.

Nevertheless, $n \to \infty$ and $h \to 0$ are not universally equivalent, because in some cases of the quantum systems, the limit $n \to \infty$ does not yield the classical one, while the limit $h \to 0$ does; the two results are not universally equivalent. The case of a particle trapped in a cubical box would be a good example: the frequency in the high quantum number domain turns out to be displaced as

$$\upsilon_q^{n+1} = \upsilon_q^n + h/2md,$$

where m is the particle's mass and d is the length of the box. Such a spectrum does not collapse toward the classical frequency in the limit of large quantum numbers, while the spectrum of the particle does degenerate to the classical continuum in the limit $h \to 0$ It can be argued that such correspondence would face another obvious problem relating to the assumption that Planck's constant goes to zero. What is the meaning of saying that "a constant goes to zero"? A constant is a number that has the same value at all times, and having it as zero is contradictory, unless it is zero. A reply to this problem might be that in correspondence, we ought to take the real limiting value and not the abstract one. In the case of relativity, the limit, "*c* goes to infinity" is an abstract one, and the real limit should be "$\upsilon/c$ goes to zero." Now, when dealing with corresponding quantum mechanics to classical mechanics, one might say that we ought to take the limit $n \to \infty$ as a better one than $h \to 0$ The point here is that values like *c* and *h* are constants and would not tend to go to zero or to infinity, but *n* and $\upsilon/c$ are variables—$n = (0, 1, 2, 3, \ldots)$ and $\upsilon/c$ varies between 0 (when $\upsilon = 0$) and 1 (when $\upsilon = c$). (Of course, this point can also count against the old CP of Planck, the first correspondence principle in our list, because it is built on the assumption that the limit is of Planck's constant going to zero.)

### *Form Correspondence Principle*

The last type of correspondence is form CP, which claims that we can obtain correspondence if the functional (mathematical) form of the new theory is the same as that of the old theory. This kind of correspondence is especially fruitful in particular cases in which other kinds of correspondence do not apply. Let us take the example used in frequency correspondence (quantum frequency). As seen in the case of the particle in a cubical box, the outcome of $n \to \infty$ does not coincide with the outcome of $h \to 0$. Hence the two limits fail to achieve the same result. In cases such as this, form correspondence might overcome the difficulties

facing frequency correspondence. The aim of form correspondence is to prove that classical frequency and quantum frequency have the same form. So, if $\upsilon_Q$ denotes quantum frequency, $\upsilon_C$ classical frequency, and $E$ energy, then form correspondence is satisfied if $\upsilon_C(E)$ has the same functional form as $\upsilon_Q(E)$. Then, by using a dipole approximation, Liboff showed that the quantum transition between state $s+n$ and state $s$ where $s \gg n$ gives the relation

$$\upsilon_Q^n(E) \approx n(E_s/2ma^2)^{1/2}. \tag{5}$$

He also noticed that if we treat the same system classically (particles of energy $E$ in a cubical box), the calculation of the radiated power in the $n$th vibrational mode is given by the expression

$$\upsilon_C^n(E) \approx n(E/2ma^2)^{1/2}. \tag{6}$$

Both frequencies have the same form, even if one is characterizing quantum frequency and the other classical, and even if their experimental treatment differs. Hence, form CP is satisfied.

But such correspondence is not problem free; in the classical case, $E$ denotes the average energy value of an ensemble of $n$th harmonic frequency, but in the quantum case, it denotes the eigenenergy of that level. Also, in the quantum case, the energy is discrete, and the only way to assert that the quantum frequency yields the classical one is by saying that when the quantum number is very big, the number of points that coincide with the classical frequency will increase, using the dipole approximation, which asserts that the distance between the points in the quantum case is assumed small. Hence the quantum case does not resemble the classical case as such, but it coincides with the average of an ensemble of classical cases.

The main thrust of form correspondence is that it can relate a branch of physics to a different branch on the basis of form resemblance, such as in the case of superconductivity. Here, a quantum formula corresponds to classical equations if we can change the quantum formula in the limit into a form where it looks similar to a classical form. The case of Josephson junctions in superconductivity, which are an important factor in building superconducting quantum interference devices, presents a perfect demonstration of such concept.

Brian David Josephson proved that the relation between the phase difference and the voltage is given by $\frac{\partial \delta}{\partial t} = \frac{2e}{\hbar}V$, that is, the voltage $V = \frac{\hbar}{2e}\frac{\partial \delta}{\partial t}$. Now, by the assertion that the Josephson junction would behave as a classical circuit, the total current would be

$$I = I_c \sin\delta + \frac{\eta}{2\mathrm{Re}}\frac{d\delta}{dt} + \frac{\eta C}{2e}\frac{d^2\delta}{dt^2}. \tag{7}$$

This equation relates the current with the phase difference but without any direct reference to the voltage. Furthermore, if we apply form correspondence, Equation 7 is analogous to the equation of a pendulum in classical mechanics. The total torque $\tau$ on the pendulum would be

$$\tau = M\frac{d^2\theta}{dt^2} + D\frac{d\theta}{dt} + \tau_0 \sin\theta, \tag{8}$$

where $M$ is the moment of inertia, $D$ is the viscous damping, and $\tau$ is the applied torque.

Both these equations have the general mathematical form

$$Y = Y_0 \sin x + B\frac{dx}{dt} + A\frac{d^2x}{dt^2}. \tag{9}$$

This kind of correspondence can be widely used to help in the solution of many problems in physics. Therefore, to find new horizons in physics, some might even think of relating some of the new theories that have not yet applied CP. Such is the case with form corresponding quantum chaos to classical chaos. The argument runs as follows: Classical chaos exists. If quantum mechanics is to be counted as a complete theory in describing nature, then it ought to have a notion that corresponds to classical chaos. That notion can be called quantum chaos. But what are the possible things that resemble chaotic behavior in quantum systems? The reply gave rise to quantum chaos. However, it turns out that a direct correspondence between the notion of chaos in quantum mechanics and that in classical mechanics does not exist.

Therefore, form correspondence would be fruitful here. Instead of corresponding quantum chaos to classical chaos, we can correspond both of them to a third entity. Classical chaos goes in a certain limit to the form $\varphi$, and quantum chaos goes to the same form at the same limit:

$$\underset{n\to\infty}{\boldsymbol{Lim}}\, classical chaos = \varphi$$
$$\boldsymbol{Lim}_{n\to\infty}\, quantum chaos = \varphi,$$

but because we have only classical and quantum theories, then the correspondence is from one to the other, as suggested by Gordon Belot and John Earman.

In addition to these four formal forms of correspondence, many other notions of correspondence might apply, such as *conceptual correspondence,* whereby new concepts ought to resemble old concepts at the limited range of applicability of such concepts. In addition, there is *observational correspondence,* which is a weak case of correspondence whereby the quantum will correspond to what is expected to be observed classically at a certain limit. *Structural correspondence* combines elements from both form correspondence and law correspondence. Hence, scientific practice might need different kinds of correspondence to achieve new relations and to relate certain domains of applicability to other domains.

## Philosophical Implications

Usually, principles in science, such as the Archimedes principle, are universally accepted. This is not the case with CP. Although CP is considered by most physicists to be a good heuristic device, it is not accepted across the board. There are two positions: The first thinks of development in science as a mere trial-and-error process; the second thinks that science is progressive and mirrors reality, and therefore new theories cannot cast away old, successful theories but merely limit the old ones to certain boundaries.

Max Born, for example, believed that scientists depend mainly on trial and error in a shattered jungle, where they do not have any signposts in science, and it is all up to them to discover new roads in science. He advised scientists to rely not on "abstract reason" but on experience. However, to accept CP means that we accept abstract reason; it also means that we do not depend on trial and error but reason from whatever accepted knowledge we have to arrive at new knowledge.

The philosophical front is much more complex. There are as many positions regarding correspondence as there are philosophers writing on the subject. But in general, realists are the defenders of the concept, whereas positivists, instrumentalists, empiricists, and antirealists are, if not opposed to the principle, then indifferent about it. Some might accept it as a useful heuristic device, but that does not give it any authoritarian power in science.

Even among realists there is more than one position. Some such as Elie Zahar claim that the CP influence ought to stem from old theory and arrive at the new through derivative power. Heinz Post is more flexible; he accepts both ways as legitimate and suggests a *generalized correspondence principle* that ought to be applied to all the developments in science. In doing so, he is replying to Thomas Kuhn's *Structure of Scientific Revolutions,* rejecting Kuhn's claim of paradigm shift and insisting on scientific continuity. In doing so, Post is also rejecting Paul Feyerabend's concept of incommensurability.

So is CP really needed? Does correspondence relate new theories to old ones, or are the new theories deduced from old theories using CP? Can old theories really be preserved? Or what, if anything, can be preserved from the old theories? What about incommensurability between the new and the old? How can we look at correspondence in light of Kuhn's concept of scientific revolution? What happens when there is a paradigm shift? All these questions are in fact related to our interpretation of CP.

CP, realists say, would help us in understanding developments in science as miracle free (the no-miracle argument). Nevertheless, by accepting CP as a principle that new theories should uphold, we in effect are trapped within the scope of old theories. This means that if our original line of reasoning was wrong and still explains the set of observations that we obtained, then the latter theory that obeys CP will resolve the problems of the old theory within a certain limit, will no doubt continue to hold the posits of the wrong theory, and will continue to abide by its accepted boundary conditions. This means that we will not be able to see where old science went wrong. In reply, conventional realists and structural realists would argue that the well-confirmed old theories are good representations of nature, and hence any new theories should resemble them at certain limits. Well, this is the heart of the matter. Even if old theories are confirmed by experimental evidence, this is not enough to claim that the abstract theory is

correct. Why? Mathematically speaking, if we have any finite set of observations, then there are many possible mathematical models that can describe this set. Hence, how can we determine that the model that was picked by the old science was the right one?

But even if we accept CP as a heuristic device, there are many ways that the concept can be applied. Each of these ways has a different set of problems for realists, and it is not possible to accept any generalized form of correspondence.

The realist position was challenged by many philosophers. Kuhn proved that during scientific revolutions the new science adopts a new paradigm in which the wordings of the old science might continue, but with different meanings. He demonstrated such a change with mass: The concept of mass in relativity is not the same as Newtonian mass. Feyerabend asserted that the changes between new science and old science make them incommensurable with each other. Hence, the realist notion of approximating new theories to old ones is going beyond the accepted limits of approximation.

The other major recent attacks on realism come from pessimistic metainduction (Larry Laudan) on one hand and new versions of empiricist arguments (Bas van Fraassen) on the other. Van Fraassen defines his position as *constructive empiricism*. Laudan relies on the history of science to claim that the realists' explanation of the successes of science does not hold. He argues that the success of theories cannot offer grounds for accepting that these theories are true (or even approximately true). He presents a list of theories that have been successful and yet are now acknowledged to be false. Hence, he concludes, depending on our previous experience with scientific revolutions, the only reasonable induction would be that it is highly probable that our current successful theories will turn out to be false. Van Fraassen claims that despite the success of theories in accounting for phenomena (their empirical adequacy), there can never be any grounds for believing any claims beyond those about what is observable. That is, we cannot say that such theories are real or that they represent nature; we can only claim that they can account for the observed phenomena.

Recent trends in realism tried to salvage realism from these attacks, but most of these trends depend on claiming that we do not need to save the old theory as a whole; we can save only the representative part. Structural realists, such as John Worrell and Elie Zahar, claim that only the mathematical structure need be saved and that CP is capable of assisting us in saving it. Philip Kitcher asserts that only presupposition posits can survive. Towfic Shomar claims that the dichotomy should be horizontal rather than vertical and that the only parts that would survive are the phenomenological models (phenomenological realism). Stathis Psillos claims that scientific theories can be divided into two parts, one consisting of the claims that contributed to successes in science (working postulates) and the other consisting of idle components.

Hans Radder, following Roy Bhaskar, thinks that progress in science is like a production line: There are inputs and outputs; hence our old knowledge of theories and observations is the input that dictates the output (our new theories). CP is important in the process; it is a good heuristic device, but it is not essential, and in many cases it does not work.

But is CP a necessary claim for all kinds of realism to account for developments in science? Some, including Shomar, do not think so. Nancy Cartwright accepts that theories are mere tools; she thinks that scientific theories are patchwork that helps in constructing models that represent different parts of nature. Some of these models depend on tools borrowed from quantum mechanics and account for phenomena related to the microscopic world; others use tools from classical mechanics and account for phenomena in the macroscopic world. There is no need to account for any connection between these models. Phenomenological realism, too, takes theories as merely tools to construct phenomenological models that are capable of representing nature. In that case, whether the fundamental theories correspond to each other to some extent or not is irrelevant. The correspondence of theories concerns realists who think that fundamental theories represent nature and approximate its blueprint.

Currently, theoretical physics is facing a deadlock; as Lee Smolin and Peter Woit have argued, the majority of theoretical physicists are running after the unification of all forces and laws of physics. They are after the theory of everything. They are convinced that science is converging toward a final theory that represents the truth about nature. They are in a way in agreement with the

realists, who hold that successive theories of "mature science" approximate the truth more and more, so science should be in quest of the final theory of the final truth.

Theoretical representation might represent the truth about nature, but we can easily imagine that we have more than one theory to depend on. Nature is complex, and in light of the richness of nature, which is reflected in scientific practice, one may be unable to accept that Albert Einstein's request for simplicity and beauty can give the correct picture of current science when complexity and diversity appear to overshadow it. The complexity of physics forces some toward a total disagreement with Einstein's dream of finding a unified theory for everything. To some, such a dream directly contradicts the accepted theoretical representations of physics. Diversity and complexity are the main characteristics of such representations.

Nonetheless, CP is an important heuristic device that can help scientists arrive at new knowledge, but scientists and philosophers should be careful as to how much of CP they want to accept. As long as they understand and accept that there is more than one version of CP and as long as they accept that not all new theories can, even in principle, revert to old theories at a certain point, then they might benefit from applying CP. One other remark of caution: Scientists and philosophers also need to accept that old theories might be wrong; the wrong mathematical form may have been picked, and if they continue to accept such a form, they will continue to uphold a false science.

*Towfic Shomar*

***See also*** Frequency Distribution; Models; Paradigm; Positivism; Theory

**Further Readings**

Fadner, W. L. (1985). Theoretical support for the generalized correspondence principle, *American Journal of Physics, 53,* 829–838.

French, S., & Kamminga, H. (Eds.). (1993). *Correspondence, invariance and heuristics: Essays in honour of Heinz Post.* Dordrecht, the Netherlands: Kluwer Academic.

Hartmann, S. (2002). On correspondence. *Studies in History & Philosophy of Modern Physics, 33B,* 79–94.

Krajewski, W. (1977). *Correspondence principle and growth in science.* Dordrecht, the Netherlands: Reidel.

Liboff, R. (1975). Bohr's correspondence principle for large quantum numbers. *Foundations of Physics, 5*(2), 271–293.

Liboff, R. (1984). The correspondence principle revisited. *Physics Today, February,* 50–55.

Makowski, A. (2006). A brief survey of various formulations of the correspondence principle. *European Journal of Physics, 27*(5), 1133–1139.

Radder, H. (1991). Heuristics and the generalized correspondence principle. *British Journal for the Philosophy of Science, 42,* 195–226.

Shomar, T. (2001). Structural realism and the correspondence principle. *Proceedings of the conference on Mulla Sadra and the world's contemporary philosophy,* Kish, Iran: Mulla Sudra Institute.

Zahar, E. (1988). *Einstein's revolution: A study in heuristics.* LaSalle, IL: Open Court.

## Covariate

Similar to an independent variable, a covariate is complementary to the dependent, or response, variable. A variable is a covariate if it is related to the dependent variable. According to this definition, any variable that is measurable and considered to have a statistical relationship with the dependent variable would qualify as a potential covariate. A covariate is thus a possible predictive or explanatory variable of the dependent variable. This may be the reason that in regression analyses, independent variables (i.e., the *regressors*) are sometimes called covariates. Used in this context, covariates are of primary interest. In most other circumstances, however, covariates are of no primary interest compared with the independent variables. They arise because the experimental or observational units are heterogeneous. When this occurs, their existence is mostly a nuisance because they may interact with the independent variables to obscure the true relationship between the dependent and the independent variables. It is in this circumstance that one needs to be aware of and make efforts to control the effect of covariates. Viewed in this context, covariates may be called by other names, such as concomitant variables, auxiliary variables, or secondary variables. This

entry discusses methods for controlling the effects of covariates and provides examples.

## Controlling Effects of Covariates

### *Research Design*

Although covariates are neither the design variable (i.e., the independent variable) nor the primary outcome (e.g., the dependent variable) in research, they are still explanatory variables that may be manipulated through experiment design so that their effect can be eliminated or minimized. Manipulation of covariates is particularly popular in controlled experiments. Many techniques can be used for this purpose. An example is to fix the covariates as constants across all experimental treatments so that their effects are exerted uniformly and can be canceled out. Another technique is through *randomization* of experimental units when assigning them to the different experimental treatments. Key advantages of randomization are (a) to control for important known and unknown factors (the control for unknown factors is especially significant) so that all covariate effects are minimized and all experimental units are statistically comparable on the mean across treatments, (b) to reduce or eliminate both intentional and unintentional human biases during the experiment, and (c) to properly evaluate error effects on the experiment because of the sound probabilistic theory that underlies the randomization. Randomization can be done to all experimental units at once or done to experimental units within a block. *Blocking* is a technique used in experimental design to further reduce the variability in experimental conditions or experimental units. Experimental units are divided into groups called blocks, and within a group, experimental units (or conditions) are assumed to be homogeneous, although they differ between groups.

However ideal, there is no guarantee that randomization eliminates all covariate effects. Even if it could remove all covariate effects, randomization may not always be feasible due to various constraints in an experiment. In most circumstances, covariates, by their nature, are not controllable through experiment designs. They are therefore not manipulated and allowed to vary naturally among experimental units across treatments. Under such circumstances, their value is often observed, together with the value of the dependent variables. The observation can be made either before, after, or during the experiment, depending on the nature of the covariates and their influence on the dependent variables. The value of a covariate may be measured prior to the administration of experimental treatments if the status of the covariate before entering into the experiment is important or if its value changes during the experiment. If the covariate is not affected by the experimental treatments, it may be measured after the experiment. The researcher, however, should be mindful that measuring a covariate after an experiment is done carries substantial risks unless there is strong evidence to support such an assumption. In the hypothetical nutrition study example given below, the initial height and weight of pupils are not covariates that can be measured after the experiment is carried out. The reason is that both height and weight are the response variables of the experiment, and they are influenced by the experimental treatments. In other circumstances, the value of the covariate is continuously monitored, along with the dependent variable, during an experiment. An example may be the yearly mean of ocean temperatures in a long-term study by R. J. Beamish and D. R. Bouillon of the relationship between the quotas of salmon fish harvested in the prior year and the number of salmon fish returned to the spawning ground of the rivers the following year, as prior research has shown that ocean temperature changes bear considerable influence on the life of salmon fish.

### *Statistical Analysis*

After the covariates are measured, a popular statistical procedure, the analysis of covariance (ANCOVA), is then used to analyze the effect of the design variables on the dependent variable by explicitly incorporating covariates into the analytical model. Assume that an experiment has $n$ design variables and $m$ covariates; a proper statistical model for the experiment would be

$$y_{ij} = \mu + t_i + \sum_{k=1}^{m} \beta_k (x_{kij} - \overline{x}_{k..}) + \varepsilon_{ij} \qquad (1)$$

where $y_{ij}$ is the $j$th measurement on the dependent variable (i.e., the primary outcome) in the $i$th treatment; $\mu$ is the overall mean; $t_i$ is the $i$th design variable (often called *treatment* in experiment design); $x_{kij}$ is the measurement on the $k$th covariate corresponding to $y_{ij}$; $\overline{x}_{k\cdot\cdot}$ is the mean of the $x_{kij}$ values; $\beta_k$ is a linear (partial) regression coefficient for the $k$th covariate, which emphasizes the relationship between $x_{ij}$ and $y_{ij}$; and $\varepsilon_{ij}$ is a random variable that follows a specific probability distribution with zero mean. An inspection of this ANCOVA model reveals that it is in fact an integration of an analysis of variance (ANOVA) model with an ANOVA of regression model. The regression part is on the covariates (recall that regressors are sometimes called covariates). A test on the hypothesis $H_0 : \beta_k = 0$ confirms or rejects the null hypothesis of no effect of covariates on the response variable. If no covariate effects exist, Equation 1 is then reduced to an ordinary ANOVA model.

Before Equation 1 can be used, though, one needs to ensure that the assumption of homogeneity of regression slopes is met. Tests must be carried out on the hypothesis that $\beta_1 = \beta_2 = \cdots = \beta_k = 0$. In practice, this is equivalent to finding no interaction between the covariates and the independent variables on the dependent variable. Without meeting this condition, tests on the adjusted means of the treatments are invalid. The reason is that when the slopes are different, the response of the treatments varies at the different levels of the covariates. Consequently, the adjusted means do not adequately describe the treatment effects, potentially resulting in misleading conclusions. If tests indeed confirm heterogeneity of slopes, alternative methods can be sought in lieu of ANCOVA, as described by Bradley Eugene Huitema. Research has repeatedly demonstrated that even randomized controlled trials can benefit from ANCOVA in uncovering the true relationship between the independent and the dependent variables.

It must be pointed out that regardless of how the covariates are handled in a study, their value, like that of the independent variables, is seldom analyzed separately because the covariate effect is of no primary interest. Instead, a detailed description of the covariates is often given to assist the reader in evaluating the results of a study. According to Fred Ramsey and Daniel Schafer, the incorporation of covariates in an ANCOVA statistical model is (a) to control for potential confounding, (b) to improve comparison across treatments, (c) to assess model adequacy, and (d) to expand the scope of inference. The last point is supported by the fact that the experiment is conducted in a more realistic environment that allows a covariate to change naturally, instead of being fixed at a few artificial levels, which may or may not be representative of the true effect of the covariate on the dependent variable.

From what has been discussed so far, it is clear that there is no fixed rule on how covariates should be dealt with in a study. Experiment control can eliminate some obvious covariates, such as ethnicity, gender, and age in a research on human subjects, but may not be feasible in all circumstances. Statistical control is convenient, but covariates need to be measured and built into a model. Some covariates may not be observable, and only so many covariates can be accommodated in a model. Omission of a key covariate could lead to severely biased results. Circumstances, therefore, dictate whether to control the effect of covariates through experiment design measures or to account for their effect in the data analysis step. A combination of experimental and statistical control is often required in certain circumstances. Regardless of what approach is taken, the ultimate goal of controlling the covariate effect is to reduce the experimental error so that the treatment effect of primary interest can be elucidated without the interference of covariates.

## Examples

To illustrate the difference between covariate and independent variables, consider an agricultural experiment on the productivity of two wheat varieties under two fertilization regimes in field conditions, where productivity is measured by tons of wheat grains produced per season per hectare (1 hectare = 2.47 acres). Although researchers can precisely control both the wheat varieties and the fertilization regimes as the primary independent variables of interest in this specific study, they are left to contend with the heterogeneity of soil fertility, that is, the natural micro variation in soil texture, soil structure, soil nutrition, soil water supply, soil aeration, and so on. These variables are natural phenomena that are beyond the control

of the researchers. They are the covariates. By themselves alone, these covariates can influence the productivity of the wheat varieties. Left unaccounted for, these factors will severely distort the experimental results in terms of wheat produced. The good news is that all these factors can be measured accurately with modern scientific instruments or methodologies. With their values known, these variables can be incorporated into an ANCOVA model to account for their effects on the experimental results.

In health research, suppose that investigators are interested in the effect of nutrition on the physical development of elementary school children between 6 and 12 years of age. More specifically, the researchers are interested in the effect of a particular commercial dietary regime that is highly promoted in television commercials. Here, physical development is measured by both height and weight increments without the implication of being obese. In pursuing their interest, the researchers choose a local boarding school with a reputation for excellent healthy nutritional programs. In this study, they want to compare this commercial dietary regime with the more traditional diets that have been offered by the school system for many years. They have done all they can to control all foreseeable potential confounding variables that might interfere with the results on the two dietary regimes. However, they are incapable of controlling the initial height and weight of the children in the study population. Randomized assignments of participating pupils to the treatment groups may help minimize the potential damage that the natural variation in initial height and weight may cause to the validity of the study, but they are not entirely sure that randomization is the right answer to the problem. These initial heights and weights are the covariates, which must be measured before (but not after) the study and accounted for in an ANCOVA model in order for the results on dietary effects to be properly interpreted.

### Final Thoughts

Covariates are explanatory variables that exist naturally within research units. What differentiates them from independent variables is that they are of no primary interest in an investigation but are nuisances that must be dealt with. Various control measures are placed on them at either the experiment design or the data analysis step to minimize the experimental error so that the treatment effects on the major outcome can be better understood. Without these measures, misleading conclusions may result, particularly when major covariates are not properly dealt with.

Regardless of how one decides to deal with the covariate effects, either by experimental control or by data analysis techniques, one must be careful not to allow the covariate to be affected by the treatments in a study. Otherwise, the covariate may interact with the treatments, making a full accounting of the covariate effect difficult or impossible.

*Shihe Fan*

***See also*** Analysis of Covariance (ANCOVA); Independent Variable; Randomization Tests

#### Further Readings

Beach, M. L., & Meier, P. (1989). Choosing covariates in the analysis of clinical trials. *Controlled Clinical Trials, 10,* S161–S175.

Beamish, R. J., & Bouillon, D. R. (1993). Pacific salmon production trends in relation to climate. *Canadian Journal of Fisheries and Aquatic* Sciences, *50,* 1002–1016.

Cox, D. R., & McCullagh. P. (1982). Some aspects of analysis of covariance. *Biometrics, 38,* 1–17.

Huitema, B. E. (1980). *The analysis of covariance and alternatives*. Toronto, Canada: Wiley.

Montgomery, D. C. (2001). *Design and analysis of experiments* (5th ed.). Toronto, Canada: Wiley.

Ramsey, F. L., & Schafer, D. W. (2002). *The statistical sleuth: A course in methods of data analysis* (2nd ed.). Pacific Grove, CA: Duxbury.

Zhang, M., Tsiatis, A. A., & Davidian, M. (2008). Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics, 64,* 707–715.

## C Parameter

*See* Guessing Parameter

# Criterion Problem

The term *criterion problem* refers to a general problem in regression analysis, especially when used for selection, when the criterion measure that is easily obtainable is not a good approximation of the actual criterion of interest. In other words, the criterion problem refers to the problem that measures of the criterion performance behaviors in which the researcher or practitioner is interested in predicting are not readily available. For example, although specific sets of performance behaviors are desirable in academic and employment settings, often the easily obtainable measures are 1st-year grade point averages (GPAs) and supervisory ratings of job performance. The criterion problem is that those easily obtainable measures (GPA and supervisory performance ratings) are not good measures of important performance behaviors. The criterion is said to be deficient if important performance behaviors are not captured in a particular criterion measure. It can also be considered contaminated if the criterion measure also assesses things that are unrelated to the performance behaviors of interest. This entry explores the criterion problem in both academic and employment settings, describes how the criterion problem can be addressed, and examines the implications for selection research.

## Criterion Problem in Academic Settings

In academic settings, there are two typical and readily available criterion variables: GPA and student retention (whether a student has remained with the university). Although each of these criteria is certainly important, most researchers agree that there is more to being a good student than simply having good grades and graduating from the university. This is the essence of the criterion problem in academic settings: the specific behaviors that persons on the admissions staff would like to predict are not captured well in the measurement of the readily available GPA and retention variables. The first step in solving the criterion problem, then, is identifying which types of behaviors are important.

Thomas Taber and Judith Hackman provided one of the earliest attempts to model the performance behaviors relating to effective performance in undergraduate college students. After surveying many university students, staff, and faculty members, they identified 17 areas of student performance and two broad categories: academic performance and nonacademic performance. The academic performance factors, such as cognitive proficiency, academic effort, and intellectual growth, are reasonably measured in the easily obtained overall GPA; these factors are also predicted fairly well by the traditional variables that predict GPA (e.g., standardized tests, prior GPA). The nonacademic performance factors, on the other hand, are not captured well in the measurement of GPA. These factors include ethical behavior, discrimination issues, and personal growth, and none of these are well predicted by traditional predictor variables.

In more recent research, Frederick Oswald and colleagues modeled undergraduate student performance for the purposes of scale construction. By examining university mission statements across a range of colleges to determine the student behaviors in which stakeholders are ultimately interested, they identified 12 performance factors that can be grouped into three broad categories: intellectual, interpersonal, and intrapersonal. Intellectual behaviors are best captured with GPA and include knowledge, interest in learning, and artistic appreciation. Interpersonal behaviors include leadership, interpersonal skills, social responsibility, and multicultural tolerance. Intrapersonal behaviors include health, ethics, perseverance, adaptability, and career orientation. However, GPA does not measure the interpersonal or intrapersonal behaviors particularly well. It is interesting to note that this research also showed that while traditional predictors (e.g., standardized tests, prior GPA) predict college GPA well, they do not predict the nonintellectual factors; noncognitive variables, such as personality and scales developed to assess these performance dimensions, are much better predictors of the nonintellectual factors.

## Criterion Problem in Employment Settings

In employment settings, the criterion problem takes a different form. One readily available criterion for many jobs is the so-called objective performance indicator. In manufacturing positions, for example, the objective performance indicator could be number of widgets produced; in sales, it could be the dollar amount of goods sold. With these measures of performance, the criterion problem is one of contamination; external factors not under the control of the employee also heavily influence the scores on these metrics. In manufacturing positions, the availability or maintenance of the equipment can have a substantial influence on the number of widgets produced; these outside influences are entirely independent of the behavior of the employee. In sales positions, season or differences in sales territory can lead to vast differences in the dollar amount of goods sold; again, these outside influences are outside the behavioral control of the employee. Such outside influences are considered criterion contamination because although they are measured with the criterion, they have nothing to do with the behaviors of interest to the organization. Unfortunately, most of these objective criterion measures suffer from contamination without possibility of better measurement for objective metrics.

Another form of the criterion problem in employment settings concerns supervisory ratings of job performance. These supervisory ratings are by far the most common form of performance evaluation in organizations. In these situations, the criterion problem concerns what, exactly, is being measured when a supervisor is asked to evaluate the performance of an employee. Although this problem has been the subject of much debate and research for the better part of a century, some recent research has shed substantial light on the issue.

John Campbell and colleagues, after doing extensive theoretical, exploratory, statistical, and empirical work in the military and the civilian workforce, developed an eight-factor taxonomy of job performance behaviors. These eight factors included job-specific task proficiency, non–job-specific task proficiency, written and oral communication, demonstration of effort, maintenance of personal discipline, facilitation of peer and team performance, supervision or leadership, and management or administrative duties. Two key elements here are of particular importance to the criterion problem. First, these are eight factors of performance *behaviors,* specifically limiting the criterion to aspects that are under the direct control of the employee and eliminating any contaminating factors. Second, these eight factors are meant to be exhaustive of all types of job performance behaviors; in other words, any type of behavior relating to an employee's performance can be grouped under one of these eight factors. Later models of performance group these eight factors under three broad factors of task performance (the core tasks relating to the job), contextual performance (interpersonal aspects of the job), and counterproductive behaviors (behaviors going against organizational policies).

## Measurement of Performance

Solutions to the criterion problem are difficult to come by. The data collection efforts that are required to overcome the criterion problem are typically time, effort, and cost intensive. It is for those reasons that the criterion problem still exists, despite the fact that theoretical models of performance are available to provide a way to assess the actual criterion behaviors of interest. For those interested in addressing the criterion problem, several broad steps must be taken. First, there must be some identification of the performance behaviors that are of interest in the particular situation. These can be identified from prior research on performance modeling in the area of interest, or they can be developed uniquely in a particular research context (e.g., through the use of job analysis). Second, a way to measure the identified performance behaviors must be developed. This can take the form of a standardized questionnaire, a set of interview questions, and so forth. Also, if performance ratings are used as a way to measure the performance behaviors, identification of the appropriate set of raters (e.g., self, peer, supervisor, faculty) for the research context would fall under this step. The third and final broad step is the actual data collection on the criterion. Often, this is the most costly step (in terms of necessary resources) because prior theoretical and empirical work can be relied on for the first two steps. To complete the

collection of criterion data, the researcher must then gather the appropriate performance data from the most appropriate source (e.g., archival data, observation, interviews, survey or questionnaire responses). Following these broad steps enables the researcher to address the criterion problem and develop a criterion variable much more closely related to the actual behaviors of interest.

## Implications for Selection Research

The criterion problem has serious implications for research into the selection of students or employees. If predictor variables are chosen on the basis of their relations with the easily obtainable criterion measure, then relying solely on that easily available criterion variable means that the selection system will select students or employees on the basis of how well they will perform on the behaviors captured in the easily measured criterion variable, but not on the other important behaviors. In other words, if the criterion is deficient, then there is a very good chance that the set of predictors in the selection system will also be deficient. Leaetta Hough and Frederick Oswald have outlined this problem in the employment domain with respect to personality variables. They showed that if careful consideration is not paid to assessing the performance domains of interest (i.e., the criterion ends up being deficient), then important predictors can be omitted from the selection system.

The criterion problem has even more severe implications for selection research in academic settings. The vast majority of the validation work has been done using GPA as the ultimate criterion of interest. Because GPA does not capture many of the inter- and intrapersonal nonintellectual performance factors that are considered important for college students, the selection systems for college admissions are also likely deficient. Although it is certainly true that admissions committees use the available information (e.g., letters of recommendation, personal statements) to try to predict these other performance factors that GPA does not assess, the extent to which the predictors relate to these other performance factors is relatively unknown.

What is clear is that unless attention is paid to the criterion problem, the resulting selection system will likely be problematic as well. When the measured criterion is deficient, the selection system will miss important predictors unless the unmeasured performance factors have the exact same determinants as the performance factors captured in the measured criterion variable. When the measured criterion is contaminated, relationships between the predictors and the criterion variable will be attenuated, as the predictors are unrelated to the contaminating factor; when a researcher must choose a small number of predictors to be included in the selection system, criterion contamination can lead to useful predictors erroneously being discarded from the final set of predictors. Finally, when a criterion variable is simply available without a clear understanding of what was measured, it is extremely difficult to choose a set of predictors.

*Matthew J. Borneman*

***See also*** Construct Validity; Criterion Validity; Criterion Variable; Dependent Variable; Selection; Threats to Validity; Validity of Research Conclusions

### Further Readings

Austin, J. T., & Villanova, P. (1992). The criterion problem: 1917–1992. *Journal of Applied Psychology, 77,* 836–874.

Campbell, J. P., Gasser, M. B., & Oswald, F. L. (1996). The substantive nature of job performance variability. In K. R. Murphy (Ed.), *Individual differences and behavior in organizations* (pp. 258–299). San Francisco: Jossey-Bass.

Campbell, J. P., McCloy, R. A., Oppler, S. H., & Sager, C. E. (1993). A theory of performance. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection* (pp. 35–70). San Francisco: Jossey-Bass.

Hough, L. M., & Oswald, F. L. (2008). Personality testing and industrial-organizational psychology: Reflections, progress, and prospects. *Industrial & Organizational Psychology: Perspectives on Science and Practice, 1,* 272–290.

Oswald, F. L., Schmitt, N., Kim, B. H., Ramsay, L. J., & Gillespie, M. A. (2004). Developing a biodata measure and situational judgment inventory as predictors of college student performance. *Journal of Applied Psychology, 89,* 187–207.

Viswesvaran, C., & Ones, D. S. (2000). Perspectives on models of job performance. *International Journal of Selection & Assessment, 8,* 216–226.
Taber, T. D., & Hackman, J. D. (1976). Dimensions of undergraduate college performance. *Journal of Applied Psychology, 61,* 546–558.

# Criterion Validity

Also known as *criterion-related validity,* or sometimes *predictive* or *concurrent validity, criterion validity* is the general term to describe how well scores on one measure (i.e., a *predictor*) predict scores on another measure of interest (i.e., the *criterion*). In other words, a particular criterion or outcome measure is of interest to the researcher; examples could include (but are not limited to) ratings of job performance, grade point average (GPA) in school, a voting outcome, or a medical diagnosis. *Criterion validity,* then, refers to the strength of the relationship between measures intended to predict the ultimate criterion of interest and the criterion measure itself. In academic settings, for example, the criterion of interest may be GPA, and the predictor being studied is the score on a standardized math test. Criterion validity, in this context, would be the strength of the relationship (e.g., the correlation coefficient) between the scores on the standardized math test and GPA.

Some care regarding the use of the term *criterion validity* needs to be employed. Typically, the term is applied to predictors, rather than criteria; researchers often refer to the "criterion validity" of a specific predictor. However, this is not meant to imply that there is only one "criterion validity" estimate for each predictor. Rather, each predictor can have different "criterion validity" estimates for many different criteria. Extending the above example, the standardized math test may have one criterion validity estimate for overall GPA, a higher criterion validity estimate for science ability, and a lower criterion validity estimate for artistic appreciation; all three are valid criteria of interest. Additionally, each of these estimates may be moderated by (i.e., have different criterion validity estimates for) situational, sample, or research design characteristics. In this entry the criterion, the research designs that assess criterion validity, effect sizes, and concerns that may arise in applied selection are discussed.

## Nature of the Criterion

Again, the term *criterion validity* typically refers to a *specific* predictor measure, often with the criterion measure assumed. Unfortunately, this introduces substantial confusion into the procedure of criterion validation. Certainly, a single predictor measure can predict an extremely wide range of criteria, as Christopher Brand has shown with general intelligence, for example. Using the same example, the criterion validity estimates for general intelligence vary quite a bit; general intelligence predicts some criteria better than others. This fact further illustrates that there is no single criterion validity estimate for a single predictor. Additionally, the relationship between one predictor measure and one criterion variable can vary depending on other variables (i.e., moderator variables), such as situational characteristics, attributes of the sample, and particularities of the research design. Issues here are highly related to the criterion problem in predictive validation studies.

## Research Design

There are four broad research designs to assess the criterion validity for a specific predictor: predictive validation, quasi-predictive validation, concurrent validation, and postdictive validation. Each of these is discussed in turn.

### *Predictive Validation*

When examining the criterion validity of a specific predictor, the researcher is often interested in selecting persons based on their scores on a predictor (or set of predictor measures) that will predict how well the people will perform on the criterion measure. In a true predictive validation design, predictor measure or measures are administered to a set of applicants, and the researchers select applicants completely randomly (i.e., without regard to their scores on the predictor measure or measures.) The correlation between the predictor measure(s) and the criterion of interest is the *index of criterion validity.* This design has the advantage of being free

from the effects of range restriction; however, it is an expensive design, and unfeasible in many situations, as stakeholders are often unwilling to forgo selecting on potentially useful predictor variables.

### *Quasi-Predictive Validation*

Like a true predictive validation design, in a quasi-predictive design, the researcher is interested in administering a predictor (or set of predictors) to the applicants in order to predict their scores on a criterion variable of interest. Unlike a true predictive design, in a quasi-predictive validation design, the researcher will select applicants based on their scores on the predictor(s). As before, the correlation between the predictor(s) and the criterion of interest is the index of criterion validity. However, in a quasi-predictive design, the correlation between the predictor and criterion will likely be smaller because of range restriction due to selection on the predictor variables. Certainly, if the researcher has a choice between a predictive and quasi-predictive design, the predictive design would be preferred because it provides a more accurate estimate of the criterion validity of the predictor(s); however, quasi-predictive designs are far more common. Although quasi-predictive designs typically suffer from range restriction problems, they have the advantage of allowing the predictors to be used for selection purposes while researchers obtain criterion validity estimates.

### *Concurrent Validation*

In a concurrent validation design, the predictor(s) of interest to the researcher are not administered to a set of applicants; rather, they are administered only to the incumbents, or people who have already been selected. The correlation between the scores on the predictors and the criterion measures for the incumbents serves as the criterion validity estimate for that predictor or set of predictors. This design has several advantages, including cost savings due to administering the predictors to fewer people and reduced time to collection of the criterion data. However, there are also some disadvantages, including the fact that criterion validity estimates are likely to be smaller as a result of range restriction (except in the rare situation when the manner in which the incumbents were selected is completely unrelated to scores on the predictor or predictors).

Another potential concern regarding concurrent validation designs is the motivation of test takers. This is a major concern for noncognitive assessments, such as personality tests, survey data, and background information. Collecting data on these types of assessments in a concurrent validation design provides an estimate of the maximum criterion validity for a given assessment. This is because incumbents, who are not motivated to alter their scores in order to be selected, are assumed to be answering honestly. However, there is some concern for intentional distortion in motivated testing sessions (i.e., when applying for a job or admittance to school), which can affect criterion validity estimates. As such, one must take care when interpreting criterion validity estimates in this type of design. If estimates under operational selection settings are of interest (i.e., when there is some motivation for distortion), then criterion validity estimates from a predictive or quasi-predictive design are of interest; however, if estimates of maximal criterion validity for the predictor(s) are of interest, then a concurrent design is appropriate.

### *Postdictive Validation*

Postdictive validation is an infrequently used design to assess criterion validity. At its basics, postdictive validation assesses the criterion variable first and then subsequently assesses the predictor variable(s). Typically, this validation design is not employed because the predictor variable(s), by definition, come temporally before the criterion variable is assessed. However, a postdictive validation design can be especially useful, if not the only alternative, when the criterion variable is rare or unethical to obtain. Such examples might include criminal activity, abuse, or medical outcomes. In rare criterion instances, it is nearly impossible to know when the outcome will occur; as such, the predictors are collected after the fact to help predict who is at risk for the particular criterion variable. In other instances when it is extremely unethical to collect data on the criterion of interest (e.g., abuse), predictor variables are collected after the fact in order to determine who might be at risk for those criterion variables. Regardless of the

reason for the postdictive design, people who met or were assessed on the criterion variable are matched with other people who were not, typically on demographic and/or other variables. The relationship between the predictor measures and the criterion variable assessed for the two groups serves as the estimate of criterion validity.

## Effect Sizes

Any discussion of criterion validity necessarily involves a discussion of effect sizes; the results of a statistical significance test are inappropriate to establish criterion validity. The question of interest in criterion validity is, To what degree are the predictor and criterion related? or How well does the measure predict scores on the criterion variable? instead of, Are the predictor and criterion related? Effect sizes address the former questions, while significance testing addresses the latter. As such, effect sizes are necessary to quantify how well the predictor and criterion are related and to provide a way to compare the criterion validity of several different predictors.

The specific effect size to be used is dependent on the research context and types of data being collected. These can include (but are not limited to) odds ratios, correlations, and standardized mean differences. For the purposes of explanation, it is assumed that there is a continuous predictor and a continuous criterion variable, making the correlation coefficient the appropriate measure of effect size. In this case, the correlation between a given predictor and a specific criterion serves as the estimate of criterion validity. Working in the effect size metric has the added benefit of permitting comparisons of criterion validity estimates for several predictors. Assuming that two predictors were collected under similar research designs and conditions and are correlated with the same criterion variable, then the predictor with the higher correlation with the criterion can be said to have greater criterion validity than the other predictor (for that particular criterion and research context). If a criterion variable measures different behaviors or was collected under different research contexts (e.g., a testing situation prone to motivated distortion vs. one without such motivation), then criterion validity estimates are not directly comparable.

### *Statistical Artifacts*

Unfortunately, several statistical artifacts can have dramatic effects on criterion validity estimates, with two of the most common being measurement error and range restriction. Both of these (in most applications) serve to lower the observed relationships from their true values. These effects are increasingly important when one is comparing the criterion validity of multiple predictors.

#### Range Restriction

Range restriction occurs when there is some mechanism that makes it more likely for people with higher scores on a variable to be selected than people with lower scores. This is common in academic or employee selection as the scores on the administered predictors (or variables related to those predictors) form the basis of who is admitted or hired. Range restriction is common in quasi-predictive designs (because predictor scores are used to select or admit people) and concurrent designs (because people are selected in a way that is related to the predictor variables of interest in the study). For example, suppose people are hired into an organization on the basis of their interview scores. The researcher administers another potential predictor of the focal criterion in a concurrent validation design. If the scores on this new predictor are correlated with scores on the interview, then range restriction will occur. True predictive validation designs are free from range restriction because either no selection occurs or selection occurs in a way uncorrelated with the predictors. In postdictive validation designs, any potential range restriction is typically controlled for in the matching scenario.

Range restriction becomes particularly problematic when the researcher is interested in comparing criterion validity estimates. This is because observed criterion validity estimates for different predictors can be differentially decreased because of range restriction. Suppose that two predictors that truly have equal criterion validity were administered to a set of applicants for a position. Because of the nature of the way they were selected, suppose that for Predictor A, 90% of the variability in predictor scores remained after people were selected, but only 50% of the variability remained for Predictor B after selection. Because

of the effects of range restriction, Predictor A would have a higher criterion validity estimate than Predictor B would, even though each had the same true criterion validity. In these cases, one should apply range restriction corrections before comparing validity coefficients. Fortunately, there are multiple formulas available to correct criterion validity estimates for range restriction, depending on the precise mechanism of range restriction.

#### Measurement Error

Unlike range restriction, the attenuation of criterion validity estimates due to unreliability occurs in all settings. Because no measure is perfectly reliable, random measurement error will serve to attenuate statistical relationships among variables. The well-known correction for attenuation serves as a way to correct observed criterion validity estimates for attenuation due to unreliability. However, some care must be taken in applications of the correction for attenuation.

In most applications, researchers are not interested in predicting scores on the criterion *measure;* rather, they are interested in predicting standing on the criterion *construct*. For example, the researcher would not be interested in predicting the supervisory ratings of a particular employee's teamwork skills, but the researcher would be interested in predicting the true nature of the teamwork skills. As such, correcting for attenuation due to measurement error in the criterion provides a way to estimate the relationship between predictor scores and the true criterion construct of interest. These corrections are extremely important when criterion reliability estimates are different in the validation for multiple predictors. If multiple predictors are correlated with the same criterion variable in the same sample, then each of these criterion validity estimates is attenuated to the same degree. However, if different samples are used, and the criterion reliability estimates are unequal in the samples, then the correction for attenuation in the criterion should be employed before making comparisons among predictors.

Corrections for attenuation in the predictor variable are appropriate only under some conditions. If the researcher is interested in a theoretical relationship between a predictor *construct* and a criterion construct, then correcting for attenuation in the predictor is warranted. However, if there is an applied interest in estimating the relationship between a particular predictor *measure* and a criterion of interest, then correcting for attenuation in the predictor is inappropriate. Although it is true that differences in predictor reliabilities can produce artifactual differences in criterion validity estimates, these differences have substantive implications in applied settings. In these instances, every effort should be made to ensure that the predictors are as reliable as possible for selection purposes.

## Concerns in Applied Selection

In applied purposes, researchers are often interested not only in the criterion validity of a specific predictor (which is indexed by the appropriate effect size) but also in predicting scores on the criterion variable of interest. For a single predictor, this is done with the equation

$$y_i = b_0 + b_1 x_i, \qquad (1)$$

where $y_i$ is the score on the criterion variable for person $i$, $x_i$ is the score on the predictor variable for person $i$, $b_0$ is the intercept for the regression model, and $b_1$ is the slope for predictor $x$. Equation 1 allows the researcher to predict scores on the criterion variable from scores on the predictor variable. This can be especially useful when a researcher wants the performance of selected employees to meet a minimum threshold.

When multiple predictors are employed, the effect size of interest is not any single bivariate correlation but the multiple correlation between a set of predictors and a single criterion of interest (which might be indexed with the multiple $R$ or $R^2$ from a regression model). In these instances, the prediction equation analogous to Equation 1 is

$$y_i = b_0 + b_{1x1i} + b_{2x2i} + \cdots + b_p x_{pi}, \qquad (2)$$

where $x_{1i}, x_{2i}, \ldots x_{pi}$ are the scores on the predictor variables 1, 2, . . . $p$ for person $i$, $b_1, b_2, \ldots b_p$ are the slopes for predictors $x_1, x_2, \ldots x_p$, and other terms are as defined earlier. Equation 2 allows the researcher to predict scores on

a criterion variable given scores on a set of $p$ predictor variables.

### *Predictive Bias*

A unique situation arises in applied selection situations because of federal guidelines requiring criterion validity evidence for predictors that show adverse impact between *protected groups*. Protected groups include (but are not limited to) ethnicity, gender, and age. Adverse impact arises when applicants from one protected group (e.g., males) are selected at a higher rate than members from another protected group (e.g., females). Oftentimes, adverse impact arises because of substantial group differences on the predictor on which applicants are being selected. In these instances, the focal predictor must be shown to exhibit criterion validity across all people being selected. However, it is also useful to examine predictive bias.

For the sake of simplicity, predictive bias will be explicated here only in the case of a single predictor, though the concepts can certainly be extended to the case of multiple predictors. In order to examine the predictive bias of a criterion validity estimate for a specific predictor, it is assumed that the variable on which bias is assessed is categorical; examples would include gender or ethnicity. The appropriate equation would be

$$y_i = b_0 + b_{1x1i} + b_{2x2i} + b_3(x_{1i} * x_{2i}), \qquad (3)$$

where $x_{1i}$ and $x_{2i}$ are the scores on the continuous predictor variable and the categorical demographic variable, respectively, for person $i$, $b_1$ is the regression coefficient for the continuous predictor, $b_2$ is the regression coefficient for the categorical predictor, $b_3$ is the regression coefficient for the interaction term, and other terms are defined as earlier. Equation 3 has substantial implications for bias in criterion validity estimates. Assuming the reference group for the categorical variable (e.g., males) is coded as 0 and the focal group (e.g., females) is coded as 1, the $b_0$ coefficient gives the intercept for the reference group, and the $b_1$ coefficient gives the regression slope for the reference group. These two coefficients form the baseline of criterion validity evidence for a given predictor. The $b_2$ coefficient and the $b_3$ coefficient give estimates of how the intercept and slope estimates, respectively, change for the focal group.

The $b_2$ and $b_3$ coefficients have strong implications for bias in criterion validity estimates. If the $b_3$ coefficient is large and positive (negative), then the slope differences (and criterion validity estimates) are substantially larger (smaller) for the focal group. However, if the $b_3$ coefficient is near zero, then the criterion validity estimates are approximately equal for the focal and reference groups. The magnitude of the $b_2$ coefficient determines (along with the magnitude of the $b_3$ coefficient) whether the criterion scores are over- or underestimated for the focal or references groups depending on their scores on the predictor variable. It is generally accepted that for predictor variables with similar levels of criterion validity, those exhibiting less predictive bias should be preferred over those exhibiting more predictive bias. However, there is some room for tradeoffs between criterion validity and predictive bias.

*Matthew J. Borneman*

***See also*** Concurrent Validity; Correction for Attenuation; Criterion Problem; Predictive Validity; Restriction of Range; Selection; Validity of Measurement

**Further Readings**

Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology, 74,* 478–494.

Brand, C. (1987). The importance of general intelligence. In S. Modgil & C. Modgil (Eds.), *Arthur Jensen: Consensus and controversy* (pp. 251–265). Philadelphia: Falmer Press.

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement, 5,* 115–124.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences.* San Francisco: W. H. Freeman.

Kuncel, N. R., & Hezlett, S. A. (2007). Standardized tests predict graduate students' success. *Science, 315,* 1080–1081.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kabin, M. B. (2001). High-stakes testing in employment, credentialing, and higher education: Prospects in

a post-affirmative action world. *American Psychologist, 56,* 302–318.
Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124,* 262–274.

# Criterion Variable

Criterion variable is a name used to describe the dependent variable in a variety of statistical modeling contexts, including multiple regression, discriminant analysis, and canonical correlation. The goal of much statistical modeling is to investigate the relationship between a (set of) criterion variable(s) and a set of predictor variables. The outcomes of such analyses are myriad and include as possibilities the development of model formulas, prediction rules, and classification rules. Criterion variables are also known under a number of other names, such as *dependent variable, response variable, predictand,* and *Y*. Similarly, predictor variables are often referred to using names such as independent variable, explanatory variable, and *X*. While such names are suggestive of a cause-and-effect relationship between the predictors and the criterion variable(s), great care should be taken in assessing causality. In general, statistical modeling alone does not establish a causal relationship between the variables but rather reveals the existence or otherwise of an observed association, where changes in the predictors are concomitant, whether causally or not, with changes in the criterion variable(s). The determination of causation typically requires further investigation, ruling out the involvement of confounding variables (other variables that affect both explanatory and criterion variables, leading to a significant association between them), and, often, scientifically explaining the process that gives rise to the causation. Causation can be particularly difficult to assess in the context of observational studies (experiments in which both explanatory and criterion variables are observed). In designed experiments, where, for example, the values of the explanatory variables might be fixed at particular, prechosen levels, it may be possible to assess causation more effectively as the research design more readily permits the adjustment of certain explanatory variables in isolation from the others, allowing a clearer judgment to be made about the nature of the relationship between the response and predictors. This entry's focus is on types of criterion variables and analysis involving criterion variables.

## Types of Criterion Variables

Criterion variables can be of several types, depending on the nature of the analysis being attempted. In many cases, the criterion variable is a measurement on a continuous or interval scale. This case is typical in observational studies, in which the criterion variable is often the variable of most interest among a large group of measured variables that might be used as predictors. In other cases, the criterion variable may be discrete, either ordinal (ordered categories) or nominal (unordered categories). A particularly important case is that of a binary (0/1) criterion variable, for which a common modeling choice is the use of logistic regression to predict the probability of each of the two possible outcomes on the basis of the values of the explanatory variables. Similarly, a categorical criterion variable requires particular modeling choices, such as multinomial logistic regression, to accommodate the form of the response variable. Increasingly, flexible nonparametric methods such as classification trees are being used to deal with categorical or binary responses, the strength of such methods being their ability to effectively model the data without the need for restrictive classical assumptions such as normality.

## Types of Analysis Involving Criterion Variables

The types of analysis that can be used to describe the behavior of criterion variables in relation to a set of explanatory variables are likewise broad. In a standard regression context, the criterion variable is modeled as a (usually) linear function of the set of explanatory variables. In its simplest form, linear regression also assumes zero-mean, additive, homoscedastic errors. *Generalized linear models* extend simple linear regression by modeling a function (called the *link function*) of the criterion variable in terms of a linear function (called the *linear predictor*) of the explanatory variables.

The criterion variable is assumed to arise from an exponential family distribution, the type of which leads to a canonical link function, the function for which $X^T Y$ is a sufficient statistic for $\beta$, the vector of regression coefficients. Common examples of exponential family distributions with their corresponding canonical link function include the *normal* (identity link), *poisson* (log link), *binomial and multinomial* (logit link), and *exponential and gamma* (inverse link) distributions. Generalized linear models allow for very flexible modeling of the criterion variable while retaining most of the advantages of simpler parametric models (compact models, easy prediction). Nonparametric models for the criterion variable include methods such as *regression trees, projection pursuit,* and *neural nets*. These methods allow for flexible models that do not rely on strict parametric assumptions, although using such models for prediction can prove challenging.

Outside the regression context, in which the goal is to model the value of a criterion variable given the values of the explanatory variables, other types of analysis in which criterion variables play a key role include *discriminant analysis,* wherein the values of the predictor (input) variables are used to assign realizations of the criterion variable into a set of predefined classes based on the values predicted for a set of linear functions of the predictors called *discriminant functions,* through a model fit via data (called the *training set*) for which the correct classes are known.

In *canonical correlation analysis,* there may be several criterion variables and several independent variables, and the goal of the analysis is to reduce the effective dimension of the data while retaining as much of the dependence structure in the data as possible. To this end, linear combinations of the criterion variables and of the independent variables are chosen to maximize the correlation between the two linear combinations. This process is then repeated with new linear combinations as long as there remains significant correlation between the respective linear combinations of criterion and independent variables. This process resembles principal components analysis, the difference being that correlation between sets of independent variables and sets of criterion variables is used as the means of choosing relevant linear combinations rather than the breakdown of variation used in principal components analysis.

Finally, in all the modeling contexts in which criterion variables are used, there exists an asymmetry in the way in which criterion variables are considered compared with the independent variables, even in observational studies, in which both sets of variables are observed or measured as opposed to fixed, as in a designed experiment. Fitting methods and measures of fit used in these contexts are therefore designed with this asymmetry in mind. For example, least squares and misclassification rates are based on deviations of realizations of the criterion variable from the predicted values from the fitted model.

*Michael A. Martin and Steven Roberts*

***See also*** Canonical Correlation Analysis; Covariate; Dependent Variable; Discriminant Analysis

**Further Readings**

Breiman, L., Friedman, J., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth.

Friedman, J. H., & Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association, 76,* 817–823.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika, 28,* 321–377.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. New York: Chapman & Hall.

McCulloch, W., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical. Biophysics, 5,* 115–133.

# Critical Difference

Critical differences can be thought of as critical regions for a priori and post hoc comparisons of pairs of means and of linear combinations of means. Critical differences can be transformed into confidence intervals. First, this entry discusses critical differences in the context of multiple comparison tests for means. Second, this entry addresses confusion surrounding applying critical differences for statistical significance and for the special case of consequential or practical significance.

## Means Model

Multiple comparison tests arise from parametric and nonparametric tests of means, medians, and ranks corresponding to different groups. The parametric case for modeling means can be described as

$y_{ij} = \mu_i + \varepsilon_{ij}$, which often assumes $\varepsilon_{ij} \overset{IID}{\sim} N(o, \sigma^2)$,

where $i = 1 \ldots p$ (number of treatments) and $j = 1 \ldots n_i$ (sample size of the $i$th treatment). The null hypothesis is that all the means are equal:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_p$$

$$H_1 : \mu_i \neq \mu_i;\ i \neq j,$$

and it is tested with an $F$ test. Regardless of the result of this test, a priori tests are always considered. However, post hoc tests are considered only if the $F$ test is significant. Both a priori and post hoc tests compare means and/or linear combinations of means. Comparing means is a natural follow-up to rejecting $H_0 : \mu_1 = \mu_2 = \ldots = \mu_p$.

## Multiple Comparison Tests for Means

Multiple comparison tests serve to uncover which pairs of means or linear contrasts of means are significant. They are often applied to analyze the results of an experiment. When the null hypothesis is that all means are equal, it is natural to compare pairs of means:

$$H_0 : \mu_i = \mu_j$$

$$H_1 : \mu_i \neq \mu_j.$$

The straightforward comparison of the two means can be generalized to a linear contrast:

$$H_0 : \sum c_i \mu_i = 0$$

$$H_1 : \sum c_i \mu_i \neq 0.$$

Linear contrasts describe testing other combinations of means. For example, the researcher might test whether the third treatment mean in an experiment is different from the average of the first and second treatment means, $\mu_3 \neq (\mu_1 + \mu_2)/2$. This can be expressed as

$$H_0 : \mu_3 - \frac{1}{2}\mu_1 - \frac{1}{2}\mu_2 = 0$$

$$H_1 : \mu_3 - \frac{1}{2}\mu_1 - \frac{1}{2}\mu_2 \neq 0.$$

Table 1 contains equations describing critical differences corresponding to typical multiple comparison tests. In addition to means, there are nonparametric critical differences for medians and for ranks. The number of treatments is denoted by $p$; the $i$th treatment group contains $n_i$ observations; $N = \sum n_i$ denotes the total number of observations in the experiment; $\alpha$ denotes the error rate for each comparison; and $k$ is the number of experimentwise comparisons—where applicable. The experimentwise error rate refers to the overall error rate for the entire experiment. Many of the critical differences in the table can be adjusted to control the experimentwise error rate.

This table is not exhaustive and is intended to be illustrative only. Many of the critical differences have variations.

These same critical differences can be used for constructing confidence intervals. However, caution is warranted as they might not be efficient.

## Consequential or Practical Significance

Multiple comparison tests are designed to find statistical significance. Sometimes the researcher's objective is to find consequential significance, which is a special case of statistical significance. Finding statistical significance between two means indicates that they are discernible at a given level of significance, $\alpha$. *Consequential significance* indicates that they are discernible and the magnitude of the difference is large enough to generate consequences. The corresponding confusion is pandemic. In order to adjust the critical differences to meet the needs of consequential significance, an extra step is wanted.

There are two classic statistical schools:

1. Fisher (Fisherian): This school does not need an alternative hypothesis.

**Table 1** Critical Difference Examples

| *Test* | *Paired Comparisons* $H_0 : \mu_i = \mu_j$ | *Critical Differences For Contrasts* $H_0 : \sum c_i \mu_i = 0$ |
|---|---|---|
| Bonferroni's | $t_{\alpha/2k,N-p}\hat{\sigma}\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}$ | $t_{\alpha/2k,N-p}\hat{\sigma}\sqrt{\sum\frac{c_i^2}{n_i}}$ |
| Duncan's method | $q(\alpha_k, p, N-p)\frac{\hat{\sigma}}{\sqrt{n}}$, where $\alpha_k = 1-(1-\alpha)^{k-1}$ | |
| Fisher's least significant difference | $t_{\alpha/2,N-p}\hat{\sigma}\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}$ | $t_{\alpha/2,N-p}\hat{\sigma}\sqrt{\sum\frac{c_i^2}{n_i}}$ |
| Multivariate $t$ method | | $t_{\alpha/2,k,N-p}\hat{\sigma}\sqrt{\sum\frac{c_i^2}{n_i}}$ |
| Newman–Keuls's $n_i = n_j$ | Comparing the order statistics of the means. $q(\alpha, p, N-p)\frac{\hat{\sigma}}{\sqrt{n}}$ | |
| Newman–Keuls's $n_i \cong n_j$ | Comparing the order statistics of the means. $q(\alpha, p, N-p)\frac{\hat{\sigma}}{\sqrt{p\left(\frac{1}{n_1}+\frac{1}{n_2}+\ldots+\frac{1}{n_t}\right)^{-1}}}$ | |
| Scheffé's | | $\sqrt{(p-1)F_{\alpha,p-1,N-2}}\hat{\sigma}\sqrt{\sum\frac{c_i^2}{n_i}}$ |
| Tukey's $n_i = n_j$ | $q(\alpha, p, N-p)\hat{\sigma}\sqrt{\frac{2}{n}}$ | $q(\alpha, p, N-p)\frac{\hat{\sigma}}{\sqrt{n}}\left(\frac{1}{2}\sum\lvert c_i\rvert\right)$ |
| Tukey's $n_i \neq n_j$ | $q(\alpha, p, N-p)\hat{\sigma}\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}; i \neq j$ | $q(\alpha, p, N-p)\frac{\hat{\sigma}}{MIN\left(\sqrt{n_i}, \sqrt{n_j}\right)}$ |

2. Neyman–Pearson: This school insists on an alternative hypothesis:

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_1 : \mu_1 - \mu_2 \neq 0$$

Ronald Fisher was the first to address the matter of hypothesis testing. He defined statistical significance to address the need for discerning between two treatments. His school emphasizes determining whether one treatment mean is greater than the other, without stressing the magnitude of the difference. This approach is in keeping with the usual critical differences as illustrated in Table 1. They pronounce whether any difference between the means is discernible and whether the ordering is statistically reliable.

J. Neyman would say that the alternative hypothesis, $H_1$, should represent the consequential scenario. For example, suppose that if the mean of Treatment 1 exceeds the mean of Treatment 2 by at least some amount $c$; then changing from Treatment 2 to Treatment 1 is consequential. The hypotheses might take the following form:

$$H_0 : \mu_1 - \mu_2 - c = 0$$
$$H_1 : \mu_1 - \mu_2 - c > 0,$$

where $c$ is the consequential difference. In the application and instruction of statistics, the two

classical schools are blended, which sows even more confusion.

To detect consequential significance, the final step consists of adjusting the multiple comparison test by adding or multiplying by constants, which might be derived from economic or scientific calculations.

*Randy J. Bartlett*

*See also* Critical Value; Margin of Error; Multiple Comparison Tests; Significance, Statistical

**Further Readings**

Fisher, R. (1925). *Statistical methods for research workers*. Edinburgh, UK: Oliver & Boyd.

Hsu, J. (1996). *Multiple comparisons, theory and methods*. Boca Raton, FL: Chapman & Hall/CRC.

Milliken, G., & Johnson, D. (1984). *Analysis of messy data: Vol. 1. Designed experiments*. Belmont, CA: Wadsworth.

Neyman, J., & Pearson, E. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika, 20A,* 175–240.

Neyman, J., & Pearson, E. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part II. *Biometrika,* 20A, 263–294.

# Critical Theory

A theory consists of a belief or beliefs that allow researchers to examine and analyze the world around them. This entry examines critical theory and how critical ideas can be applied to research. As Peter Barry suggests, we can use theory rather than theory's using us. The complexity of any theory, let alone the critical theories examined in this entry, cannot be underestimated, but theory can be used to produce better research methodology and also to test research data. Theory is important because it allows us to question our beliefs, meanings, and understandings of the world around us. In all subject areas, critical theory provides a multidimensional and wide-ranging critique of research problems that researchers attempt to address. The importance of how we read, apply, and interpret critical theories is crucial within any research design. Within this entry, the aim is to increase understandings of Marxism, critical race theory, postmodernism, and poststructuralism and also to show, through the evidence bases of literature and reflective experiences, how critical theories can be used by the researcher within different parts of research design.

## Marxist Critique

Marx's critique of the free market economy and capitalism is as applicable today as it was in the 19th century. Capitalism as an ideology was a power structure based on exploitation of the working classes. Economic production revolves around the exploited relationship between bourgeoisie and proletariat. The bourgeoisie have monopolized production since the beginning of the industrial revolution, forcing the proletariat from the land and into structured environments (e.g., urban factories). Peasants became workers, and the working wage was a controlled, structured device that gives the proletariat only a fraction of the generated revenue. The surplus value from production is a profit that is taken by the bourgeoisie. Critical theorists would argue that this practice is both theft and wage slavery. This leads to wider consequences (i.e., overproduction), which in the 20th and 21st centuries has had wide implications for the environment. Marxism favors social organization with all individuals having the right to participate in consumption and production. The result of class exploitation, Marxists believe, would be the revolutionary overthrow of capitalism with communism.

Max Horkheimer, a founding member of the Frankfurt School in the 1920s and 1930s, wanted to develop critical theory as a form of Marxism, with the objective of changing society. His ideas attempted to address the threat not only of fascism in Europe but of the rise of consumer culture, which can create forms of cultural hegemony and new forms of social control. Theory adapts to changing times, but within new social situations, theories can be developed and applied to contemporary contexts. Horkheimer's focus on consumer culture can be used today within modern communication when we reflect on how people today socialize. One example is Facebook and the ways behavioral patterns continue to change as people use the World Wide Web more and more to exchange personal information. What are the

consequences of these changes for social organization and consumer culture?

A very contemporary application of Marxism has been provided by Mike Cole, who applies Marxist ideas to education by using the example of Venezuela and Hugo Chavez, who opposes capitalism and imperialism. In education, Cole highlights, Chavez has hoped to open 38 new state universities with 190 satellite classrooms throughout Venezuela by 2009. Social projects such as housing are linked to this policy. Communal councils have been created whereby the local population meets to decide on local policies and how to implement them, rather than relying on bourgeois administrative machinery. Chavez is not only talking about democratic socialism but applying it to government policy. Therefore, one can apply Marxist critique politically in different parts of the world. Chavez's policies are a reaction to capitalism and the colonial legacy and have the objective of moving Venezuela in a more socialist direction. Application and interpretation are the keys when applying critical theory within research design. Cole applies Marxist ideas to the example of Venezuela and provides evidence to interpret the events that are taking place. Critical theory can be effective when it is applied in contemporary contexts.

## Critical Race Theory

The critical race theory (CRT) movement is a collection of activists and scholars interested in studying and transforming the relationship among race, racism, and power. Although CRT began as a movement in the United States within the subject of law, it has rapidly spread beyond that discipline. Today, academics in the social and behavioral sciences, including the field of education, consider themselves critical race theorists who use CRT ideas to, according to Richard Delgardo and Jean Stefancic, understand issues of school discipline, controversies, tracking, and IQ and achievement testing. CRT tries not only to understand our social situation but to change it. The focus of CRT is racism and how it is socially and culturally constructed. CRT goes beyond a conceptual focus of multiculturalism, which examines equal opportunities, equity, and cultural diversity. Barry Troyna carried out education research in the United Kingdom during the 1970s and 1980s with an antiracist conceptual analysis, arguing that multiculturalism did not focus on antiracist practice enough. CRT has that central focus and goes further in relation to how minority groups are racialized and colored voices silenced. CRT offers everyone a voice to explore, examine, debate, and increase understandings of racism. David Gillborn is an advocate of applying CRT within research, explaining that the focus of CRT is an understanding that the status of Black and other minority groups is always conditional on Whites. As Gillborn highlights, to many Whites, such an analysis might seem outrageous, but its perceptiveness was revealed in dramatic fashion in July 2005, with the terrorist attacks on London and Madrid. Gillborn underlines that this was the clearest demonstration of the conditional status of people of color in contemporary England. So power conditions are highlighted by CRT. The application of CRT within research design can be applied in many subject areas and disciplines. That notion of power and how it is created, reinforced, and controlled is an important theme within critical theory.

## Postmodernism

Postmodernism also examines power relations and how power is made and reinforced. To understand what postmodernism is, we have to understand what modernism was. Peter Barry explains that modernism is the name given to the movement that dominated arts and culture in the first half of the 20th century. Practice in music, literature, and architecture was challenged. This movement of redefining modernism as postmodernist can be seen in the changing characteristics of literary modernism. Barry provides a position that concerns literary modernism and the move to postmodernist forms of literature. It can be applied more broadly to research design and the application of critical theory. The move away from grand-narrative social and cultural theories can be seen in the above as philosophers and cultural commentators moved away from "objective" positions and began to examine multiple points of views and diverse moral positions. Postmodernists are skeptical about overall answers to questions that allow no space for debate. Reflexivity allows the researcher to reflect on his or her own identity or identities within a given profession. The idea of moving beyond a simplistic mirror image or "Dear Diary"

approach to a reflective method with the application of different evidence bases of literature reviews to personal experiences shows how status, roles, and power can be critically analyzed. The plurality of roles that occurs is also a postmodern development with the critical questioning of the world and the multiple identities that globalization gives the individual. In relation to research design, a postmodern approach gives the researcher more possibilities in attempting to increase understandings of a research area, question, or hypothesis. Fragmented forms, discontinuous narrative, and the random nature of material can give the researcher more areas or issues to examine. This can be problematic as research focus is an important issue in the research process and it is vital for the researcher to stay focused. That last line would immediately be questioned by a postmodernist because the position is one of constant questioning and potential change. The very nature of research and research design could be questioned by the postmodernist. The issue here is the creation and development of ideas, as it is for all researchers. Jean-François Lyotard believed that the researcher and intellectual should resist the grand ideas and narratives that had become, in his opinion, outdated. Applying that directly to research design, a modernist argument would be that the research process should consist of an introduction with a research question or hypothesis; literature review; method and methodology; data collection, presentation, and analysis; and a conclusion. If one were being critical of that provisional research design structure, one could suggest that research questions (plural) should be asked, literature reviews (subject specific, general, theoretical, conceptual, method, data) should be carried out, and all literature should be criticized; positivist research paradigms should be dropped in favor of more reflective, action research projects; and other questions should be answered rather than the focal question or hypothesis posed at the beginning of the research project. Research design itself would be questioned because that is the very nature of postmodernist thought: the continuing critique of the subject under examination.

## Poststructuralism

The issue of power and knowledge creation is also examined within poststructuralism in the sense that this form of critical theory aims to deconstruct the grand narratives and structural theoretical frameworks. Poststructuralism also attempts to increase understandings of language and the ways knowledge and power are used and evolve to shape how we view structures (e.g., the institution in which we work and how it works) and why we and others accept how these structures work. Poststructuralism is critical of these processes and attempts to analyze new and alternative meanings. In relation to research design, it is not only how we apply critical theory to poststructuralist contexts, it is how we attempt to read the theory and theorists. Michel Foucault is a poststructuralist, and his works are useful to read in association with issues of research design. Poststructural ideas can be used to examine the meanings of different words and how different people hold different views or meanings of those words. The plurality of poststructuralist debate has been criticized because considering different or all arguments is only relative when an absolute decision has to be taken. However, it is the question of language and meaning in relation to power and knowledge that offers the researcher a different angle within research design.

## Application Within the Social and Behavioral Sciences

This final section highlights how critical theory, be it Marxism, CRT, postmodernism, or poststructuralism, can be applied within the social and behavioral sciences. It is not only the word *application* that needs to be focused on but interpretation and one's interpretations in relation to one's research question or hypothesis. Researchers reading the primary sources of Marx, Foucault, or Lyotard and applying them to a research design is all very well, but interpreting one's own contextual meanings to methodology from the literature reviews and then applying and interpreting again to data analysis seems to be more difficult. Two different meanings could materialize here, which is due to the fact that space and time needs to be given within research design for reading and rereading critical theories to increase understandings of what is being researched. Critical theories can be described as windows of opportunity in exploring research processes in the social and behavioral sciences. They

can be used as a tool within research design to inform, examine, and ultimately test a research question and hypothesis. Critical theoretical frameworks can be used within research introductions, literature reviews, and method and methodology processes. They can have a role to play in data analysis and research conclusions or recommendations at the end of a research project. It is how the researcher uses, applies, and interprets critical theory within research design that is the key issue.

*Richard Race*

***See also*** Literature Review; Methods Section; Research Design Principles; Research Question; Theory

**Further Readings**

Delgardo, R., & Stefancic, J. (2001). *Critical race theory: An introduction.* New York: New York University Press.
Foucault, M. (1991). *Discipline and punish: The birth of the prison.* London: Penguin Books.
Foucault, M. (2002). *Archaeology of knowledge.* London: Routledge.
Lyotard, J.-F. (1984). *The postmodern condition: A report on knowledge.* Minneapolis: University of Minnesota Press.
Malpas, S., & Wake, P. (Eds.). (2006). *The Routledge companion to critical theory.* London: Routledge.
Troyna, B. (1993). *Racism and education.* Buckingham, UK: Open University Press.

# Critical Thinking

Critical thinking evaluates the validity of propositions. It is the hallmark and the cornerstone of science because science is a community that aims to generate true statements about reality. The goals of science can be achieved only by engaging in an evaluation of statements purporting to be true, weeding out the false ones, and limiting the true ones to their proper contexts. Its centrality to the scientific enterprise can be observed in the privileges accorded to critical thinking in scientific discourse. It usually trumps all other considerations, including tact, when it appears in a venue that considers itself to be scientific.

A *proposition* is a statement that claims to be true, a statement that claims to be a good guide to reality. Not all statements that sound as if they may be true or false function as propositions, so the first step in critical thinking is often to consider whether a proposition is really being advanced. For example, "I knew this was going to happen" is often an effort to save face or to feel some control over an unfortunate event rather than an assertion of foreknowledge, even though it sounds like one. Conversely, a statement may not sound as if it has a truth element, but on inspection, one may be discovered. "Read Shakespeare" may sometimes be translated as the proposition, "Private events are hard to observe directly, so one way to learn more about humans is to observe public representations of private thoughts as described in context by celebrated writers." Critical thinking must evaluate statements properly stated as propositions; many disagreements are settled simply by ascertaining what, if anything, is being proposed. In research, it is useful to state hypotheses explicitly and to define the terms of the hypotheses in a way that allows all parties to the conversation to understand exactly what is being claimed.

Critical thinking contextualizes propositions; it helps the thinker consider when a proposition is true or false, not just whether it is true or false. If a proposition is always true, then it is either a tautology or a natural law. A *tautology* is a statement that is true by definition: "All ermines are white" is a tautology in places where nonwhite ermines are called weasels. A natural *law* is a proposition that is true in all situations, such as the impossibility of traveling faster than light in a vacuum. The validity of all other propositions depends on the situation. Critical thinking qualifies this validity by specifying the conditions under which they are good guides to reality.

*Logic* is a method of deriving true statements from other true statements. A *fallacy* occurs when a false statement is derived from a true statement.

This entry discusses methods for examining propositions and describes the obstacles to critical thinking.

## Seven Questions

Critical thinking takes forms that have proven effective in evaluating the validity of propositions.

Generally, critical thinkers ask, in one form or another, the following seven questions:

1. What does the statement assert? What is asserted by implication?
2. What constitutes evidence for or against the proposition?
3. What is the evidence for the proposition? What is the evidence against it?
4. What other explanations might there be for the evidence?
5. To which circumstances does the proposition apply?
6. Are the circumstances currently of interest like the circumstances to which the proposition applies?
7. What motives might the proponent of the proposition have besides validity?

### *What Does the Statement Assert? What Is Asserted by Implication?*

The proposition *Small schools produce better citizens than large schools do* can be examined as an illustrative example. The first step requires the critical thinker to define the terms of the proposition. In this example, the word *better* needs elaboration, but it is also unclear what is meant by *citizen.* Thus, the proponent may mean that better citizens are those who commit fewer crimes or perhaps those who are on friendly terms with a larger proportion of their communities than most citizens.

Critical thinkers are alert to hidden tautologies, or to avoiding the fallacy of *begging the question,* in which *begging* is a synonym for *pleading* (as in pleading the facts in a legal argument) and *question* means the proposition at stake. It is fallacious to prove something by assuming it. In this example, students at smaller schools are bound to be on speaking terms with a higher proportion of members of the school community than are students at larger schools, so if that is the definition of better citizenship, the proposition can be discarded as trivial. Some questions at stake are so thoroughly embedded in their premises that only very deep critical thinking, called *deconstruction,* can reveal them. Deconstruction asks about implied assumptions of the proposition, especially about unspoken dualities. Thus, a critical thinker would want to examine whether it makes sense to consider one citizen better than another or whether the proposition is implying that schools are responsible for social conduct rather than for academics.

Critical thinkers are also alert to artificial categories. When categories are implied by a proposition, they need to be examined as to whether they really exist. Most people would accept the reality of the category *school* in the contemporary United States, but not all societies have clearly demarcated mandatory institutions where children are sent during the day. It is far from clear that the categories of smaller schools and larger schools stand up to scrutiny, because school populations, though not falling on a smooth curve, are more linear than categorical. The proponent might switch the proposition to *School size predicts later criminal activity.*

### *What Constitutes Evidence for or Against the Proposition?*

Before evidence is evaluated for its effect on validity, it must be challenged by questions that ask whether it is good evidence of anything. This is generally what is meant by reliability. If a study examines a random sample of graduates' criminal records, critical thinkers will ask whether the sample is truly random, whether the available criminal records are accurate and comprehensive, whether the same results would be obtained if the same records were examined on different days by different researchers, and whether the results were correctly transcribed to the research protocols.

It is often said that science relies on evidence rather than on ipse dixits, which are propositions accepted solely on the authority of the speaker. This is a mistaken view, because all propositions ultimately rest on ipse dixit. In tracking down criminal records, for example, researchers will eventually take someone's—or a computer's—word for something.

### *What Is the Evidence for the Proposition? What Is the Evidence Against It?*

These questions are useful only if they are asked, but frequently people ask only about the evidence on the side they are predisposed to believe. When they do remember to ask, people

have a natural tendency, called *confirmation bias,* to value the confirming evidence and to dismiss the contradictory evidence.

Once evidence is adduced for a proposition, one must consider whether the very same evidence may stand against it. For example, once someone has argued that distress in a child at seeing a parent during a stay in foster care is a sign of a bad relationship, it is difficult to use the same distress as a sign that it is a good relationship. But critical thinking requires questioning what the evidence is evidence of.

### *What Other Explanations Might There Be for the Evidence?*

To ensure that an assertion of causality is correct, either one must be able to change all but the causal variable and produce the same result, or one must be able to change only the proposed cause and produce a different result. In practice, especially in the social areas of science, this never happens, because it is extremely difficult to change only one variable and impossible to change all variables except one. Critical thinkers identify variables that changed along with the one under consideration. For example, it is hard to find schools of different sizes that also do not involve communities with different amounts of wealth, social upheaval, or employment opportunities. Smaller schools may more likely be private schools, which implies greater responsiveness to the families paying the salaries, and it may be that responsiveness and accountability are more important than size per se.

### *To Which Circumstances Does the Proposition Apply?*

If a proposition is accepted as valid, then it is either a law of nature—always true—or else it is true only under certain circumstances. Critical thinkers are careful to specify these circumstances so that the proposition does not become overly generalized. Thus, even if a causal relationship were accepted between school size and future criminality, the applicability of the proposition might have to be constricted to inner cities or to suburbs, or to poor or rich schools, or to schools where entering test scores were above or below a certain range.

### *Are the Circumstances Currently of Interest Like the Circumstances to Which the Proposition Applies?*

Once a proposition has been validated for a particular set of circumstances, the critical thinker examines the current situation to determine whether it is sufficiently like the validating circumstances to apply the proposition to it. In the social sciences, there are always aspects of the present case that make it different from the validating circumstances. Whether these aspects are different enough to invalidate the proposition is a matter of judgment. For example, the proposition relating school size to future criminality could have been validated in California, but it is unclear whether it can be applied to Texas. Critical thinkers form opinions about the similarity or differences between situations after considering reasons to think the current case is different from or similar to the typical case.

### *What Motives Might the Proponent of the Proposition Have Besides Validity?*

The scientific community often prides itself on considering the content of an argument rather than its source, purporting to disdain ad hominem arguments (those being arguments against the proponent rather than against the proposition). However, once it is understood that all evidence ultimately rests on ipse dixits, it becomes relevant to understand the motivations of the proponent. Also, critical thinkers budget their time to examine relevant propositions, so a shocking idea from a novice or an amateur or, especially, an interested party is not always worth examining. Thus, if a superintendent asserts that large schools lead to greater criminality, one would want to know what this official's budgetary stake was in the argument. Also, this step leads the critical thinker full circle, back to the question of what is actually being asserted. If a high school student says that large schools increase criminality, he may really be asking to transfer to a small school, a request that may not depend on the validity of the proposition.

Thus, critical thinkers ask who is making the assertion, what is at stake for the speaker, from what position the speaker is speaking, and under what conditions or constraints, to what audience, and with what kind of language. There may be no

clear or definitive answers to these questions; however, the process of asking is one that actively engages the thinking person in the evaluation of any proposition as a communication.

## The Role of Theory

When critical thinkers question what constitutes good evidence, or whether the current situation is like or unlike the validating circumstances, or what motives the proponent may have, how do they know which factors to consider? When they ask about other explanations, where do other explanations come from? Theory, in the sense of a narrative that describes reality, provides these factors and explanations. To use any theorist's theory to address any of these issues is to ask what the theorist would say about them.

## Multiculturalism

Multiculturalism provides another set of questions to examine the validity of and especially to constrict the application of a proposition. Someone thinking of large suburban high schools and small suburban parochial schools may think that the proposition relating school size to criminality stands apart from race, sex, and ethnicity. Multicultural awareness reminds us to ask whether these factors matter.

## Obstacles to Critical Thinking

If critical thinking is such a useful process for getting at the truth, for producing knowledge that may more efficiently and productively guide our behavior—if it is superior to unquestioning acceptance of popular precepts, common sense, gut instinct, religious faith, or folk wisdom—then why is it not more widespread? Why do some people resist or reject critical thinking? There are several reasons that it can be upsetting. Critical thinkers confront at least six obstacles: losing face, losing faith, losing friends, thinking the unthinkable, challenging beliefs, and challenging believing.

### *Losing Face*

Critical thinking can cause people to lose face. In nonscientific communities, that is, in communities not devoted to generating true statements about reality, propositions are typically met with a certain amount of tact. It can be awkward to question someone's assertions about reality, and downright rude to challenge their assertions about themselves. When people say something is true and it turns out not to be true, or not to be always true, they lose face. Scientists try to overcome this loss of face by providing a method of saving face, namely, by making a virtue of self-correction and putting truth ahead of pride. But without a commitment to science's values, critical thinking can lead to hurt feelings.

### *Losing Faith*

Religious faith is often expressed in spiritual and moral terms, but sometimes it is also expressed in factual terms—faith that certain events happened at a certain time or that the laws of nature are sometimes transcended. When religion takes a factual turn, critical thinking can oppose it, and people can feel torn between religion and science. Galileo said that faith should concern itself with how to go to heaven, and not with how the heavens go. When people have faith in how reality works, critical thinking can become the adversary of faith.

### *Losing Friends*

Human beings are social; we live together according to unspoken and spoken agreements, and our social networks frequently become communities of practice wherein we express our experiences and observations in terms that are agreeable to and accepted by our friends and immediate communities. Among our intimates and within our social hierarchies, we validate each other's views of reality and find such validation comforting. We embed ourselves in like-minded communities, where views that challenge our own are rarely advanced, and when they are, they and their proponents are marginalized—labeled silly, dangerous, crazy, or unreasonable—but not seriously considered. This response to unfamiliar or disquieting propositions strengthens our status as members of the ingroup and differentiates us from the outgroup. Like-minded communities provide reassurance, and critical thinking—which looks with a curious, analytical, and fearless eye on propositions—threatens not only one's worldview but one's social ties.

### *Thinking the Unthinkable*

Some of the questions that critical thinkers ask in order to evaluate propositions require an imaginative wondering about alternatives: Would the meaning of evidence change if something about the context changed, could there be some other, unthought-of factor accounting for evidence, or could the evidence itself be suspect? Critical thinking involves considering alternatives that may be banned from discourse because they challenge the status quo or because they challenge the complacency of the thinker. It may not be tolerable to ask whether race matters, for example, in trying to understand any links between school size and criminality, and it may not be tolerable to ask how a group of professionals would behave if arrested, if one is evaluating the proposition that a suspect's behavior on arrest signifies guilt or innocence. Critical thinking can be frightening to people who just want to be left alone and do not want to think about alternatives.

### *Challenging Beliefs*

It is useful for most people to have several central tenets about reality and the human condition that guide their behavior. If anxiety is the feeling of not knowing what to do, it can be comforting to have some beliefs that dictate how to behave. Very few beliefs emerge from the process of critical thinking unchanged and unqualified. The basic propositions on which an individual relies come to seem temporary and situational rather than fundamental and reliable. The result can be an anxious sense that the map being used to navigate reality is out of date or drawn incorrectly. To preserve the security of knowing what to do, many people avoid evaluating their maps and beliefs.

### *Challenging Believing*

The very process of examining beliefs can make them seem arbitrary, like tentative propositions reflecting communicative circumstances and personal motives rather than truisms about reality etched in stone. This happens because critical thinking casts beliefs as statements, as sentences uttered or written by a speaker, and the process of questioning them makes it clear that there are almost always situational and motivational issues to consider. Critical thinking shows us the extent to which our understandings are socially constructed and communicative. With very few exceptions, science's propositions are continually revised and circumscribed, facts blur with opinions, and our categories turn out to be arbitrary or even in the service of some social agenda. A postmodern world, energized by critical thinking about the keystones of our beliefs, can be a difficult world to live in for people seeking certainty.

It is the nature of powerful subsystems to preserve their power by setting their definitions of situations in stone, and it is the task of critical thinkers to cast those definitions as revisable, circumscribable, and oftentimes arbitrary. If truth liberates people from other people's self-serving views of reality, critical thinking is an engine for freedom as well as for truth.

*Michael Karson and Janna Goodwin*

***See also*** Reliability; Theory; Threats to Validity; "Validity"

**Further Readings**

Derrida, J. (1982). *Margins of philosophy* (A. Bass, Trans.). Chicago: University of Chicago Press.

Karson, M., & Goodwin, J. (2008). Critical thinking about critical thinking. In M. Karson, *Deadly therapy: Lessons in liveliness from theater and performance theory* (pp. 159–175). Lanham, MD: Jason Aronson.

Longino, H. (1990). *Science as social knowledge: Values and objectivity in scientific inquiry.* Princeton, NJ: Princeton University Press.

Meehl, P. (1973). Why I do not attend case conferences. In P. E. Meehl, *Psychodiagnosis: Selected papers* (pp. 225–302). Minneapolis: University of Minnesota Press.

## Critical Value

The critical value (CV) is used in significance testing to establish the critical and noncritical regions of a distribution. If the test value or statistic falls within the range of the *critical region* (the *region of rejection*), then there is a significant difference or association; thus, the null hypothesis should be rejected. Conversely, if the test value or statistic falls within the range of the *noncritical region* (also known as the *nonrejection region*), then the difference or association is possibly due to chance; thus, the null hypothesis should be accepted.

When using a one-tailed test (either left-tailed or right-tailed), the CV will be on either the left or the right side of the mean. Whether the CV is on the left or right side of the mean is dependent on the conditions of an alternative hypothesis. For example, a scientist might be interested in increasing the average life span of a fruit fly; therefore, the alternative hypothesis might be $H_1 : \mu > 40$ days. Subsequently, the CV is on the right side of the mean. Likewise, the null hypothesis would be rejected only if the sample mean is greater than 40 days. This example would be referred to as a one-tailed right test.

To use the CV to determine the significance of a statistic, the researcher must state the null and alternative hypotheses; set the level of significance, or alpha level, at which the null hypothesis will be rejected; and compute the test value (and the corresponding degrees of freedom, or *df,* if necessary). The investigator can then use that information to select the CV from a table (or calculation) for the appropriate test and compare it to the statistic. The statistical test the researcher chooses to use (e.g., *z*-score test, *z* test, single sample *t* test, independent samples *t* test, dependent samples *t* test, one-way analysis of variance, Pearson product-moment correlation coefficient, chi-square) determines which table he or she will reference to obtain the appropriate CV (e.g., *z*-distribution table, *t*-distribution table, *F*-distribution table, Pearson's table, chi-square distribution table). These tables are often included in the appendixes of introductory statistics textbooks.

For the following examples, the Pearson's table, which gives the CVs for determining whether a Pearson product-moment correlation (*r*) is statistically significant, is used. Using an alpha level of .05 for a two-tailed test, with a sample size of 12 ($df = 10$), the CV is .576. In other words, for a correlation to be statistically significant at the .05 significance level using a two-tailed test for a sample size of 12, then the absolute value of Pearson's *r* must be greater than or equal to .576. Using a significance level of .05 for a one-tailed test, with a sample size of 12 ($df = 10$), the CV is .497. Thus, for a correlation to be statistically significant at the .05 level using a one-tailed test for a sample size of 12, then the absolute value of Pearson's *r* must be greater than or equal to .497.

When using a statistical table to reference a CV, it is sometimes necessary to interpolate, or estimate values, between CVs in a table because such tables are not exhaustive lists of CVs. For the following example, the *t*-distribution table is used. Assume that we want to find the critical *t* value that corresponds to 42 *df* using a significance level or alpha of .05 for a two-tailed test. The table has CVs only for 40 *df* (CV = 2.021) and 50 *df* (CV = 2.009). In order to calculate the desired CV, we must first find the distance between the two known *df*s (50 − 40 = 10). Then we find the distance between the desired *df* and the lower known *df* (42 − 40 = 2). Next, we calculate the proportion of the distance that the desired *df* falls from the lower known $df\left(\frac{2}{10} = .20\right)$. Then we find the distance between the CVs for 40 *df* and 50 $df(2.021 - 2.009 = .012)$. The desired CV is .20 of the distance between 2.021 and 2.009 $(.20 \times .012 = .0024)$. Since the CVs decrease as the *df*s increase, we subtract .0024 from CV for 40 $df(2.021 - .0024 = 2.0186)$; therefore, the CV for 42 *df* with an alpha of .05 for a two-tailed test is $t = 2.0186$.

Typically individuals do not need to reference statistical tables because statistical software packages, such as SPSS, an IBM company, formerly called PASW® Statistics, indicate in the output whether a test value is significant and the level of significance. Furthermore, the computer calculations are more accurate and precise than the information presented in statistical tables.

*Michelle J. Boyd*

***See also*** Alternative Hypotheses; Degrees of Freedom; Null Hypothesis; One-Tailed Test; Significance, Statistical; Significance Level, Concept of; Two-Tailed Test

**Further Readings**

Heiman, G. W. (2003). *Basic statistics for the behavioral sciences*. Boston: Houghton Mifflin.

## Cronbach's Alpha

***See*** Coefficient Alpha

# Crossover Design

There are many different types of experimental designs for different study scenarios. Crossover design is a special design in which each experimental unit receives a sequence of experimental treatments. In practice, it is not necessary that all permutations of all treatments be used. Researchers also call it *switchover design,* compared with a *parallel group design,* in which some experimental units only get a specific treatment and other experimental units get another treatment. In fact, the crossover design is a specific type of repeated measures experimental design. In the traditional repeated measures experiment, the experimental units, which are applied to one treatment (or one treatment combination) throughout the whole experiment, are measured more than one time, resulting in correlations between the measurements. The difference between crossover design and traditional repeated measures design is that in crossover design, the treatment applied to an experimental unit for a specific time continues until the experimental unit receives all treatments. Some experimental units may be given the same treatment in two or more successive periods, according to the needs of the research.

The following example illustrates the display of the crossover design. Researchers altered the diet ingredients of 18 steers in order to study the digestibility of feedstuffs in beef cattle. There were three treatments, or feed mixes, each with a different mix of alfalfa, straw, and so on. A three-period treatment was used, with 3 beef steers assigned to each of the six treatment sequences. Each diet in each sequence was fed for 30 days. There was a 21-day washout between each treatment period of the study. Assume that the dependent variable is the neutral detergent fiber digestion coefficient calculated for each steer. For this case, there are three treatment periods and $3! = 6$ different sequences. The basic layout may be displayed as in Table 1.

**Table 1** Six-Order Sequence Crossover Table

| *Sequence* | *Units* | *Period 1* | *Period 2* | *Period 3* |
|---|---|---|---|---|
| 1 | 1,2,3 | A | B | C |
| 2 | 4,5,6 | B | C | A |
| 3 | 7,8,9 | C | A | B |
| 4 | 10,11,12 | A | C | B |
| 5 | 13,14,15 | B | A | C |
| 6 | 16,17,18 | C | B | A |

## Basic Conceptions, Advantages, and Disadvantages

Crossover designs were first used in agriculture research in 1950s and are widely used in other scientific fields with human beings and animals. The feature that measurements are obtained on different treatments from each experimental unit distinguishes the crossover design from other experimental designs.

This feature entails several advantages and disadvantages. One advantage is reduced costs and resources when money and the number of experimental units available for the study are limited. The main advantage of crossover design is that the treatments are compared within subjects. The crossover design is able to remove the subject effect from the comparison. That is, the crossover design removes any component from the treatment comparisons that is related to the differences between the subjects. In clinical trials, it is common that the variability of measurements obtained from different subjects is much greater than the variability of repeated measurement obtained from the same subjects.

A disadvantage of crossover design is that it may bring a *carryover effect,* that is, the effect of a treatment given in one period may influence the effect of the treatment in the following period(s). Typically the subjects (experimental units) are given sufficient time to "wash out" the effect of the treatments between two periods in crossover designs and return to their original state. It is important in a crossover study that the underlying condition does not change over time and that the effects of one treatment disappear before the next is applied. In practice, even if sufficient washout time is administered after two successive treatment periods, the subjects' physiological states may have been changed and may be unable to return to the original state, which may affect the effects of the treatment in the succeeding period. Thus, the carryover effect cannot be ignored in crossover designs. In spite of its advantages, the crossover

design should not be used in clinical trials in which a treatment cures a disease and no underlying condition remains for the next treatment period. Crossover designs are typically used for persistent conditions that are unlikely to change over the course of the study. The carryover effect can cause problems with data analysis and interpretation of results in a crossover design. The carryover prevents the investigators from determining whether the significant effect is truly due to a direct treatment effect or whether it is a residual effect of other treatments. In multiple regressions, the carryover effect often leads to multicollinearity, which leads to erroneous interpretation. If the crossover design is used and a carryover effect exists, a design should be used in which the carryover effect will not be confounded with the period and treatment effects. The carryover effect makes the design less efficient and more time-consuming.

The *period effect* occurs in crossover design because of the conditions that are present at the time the observed values are taken. These conditions systematically affect all responses that are taken during that time, regardless of the treatment or the subject. For example, the subjects need to spend several hours to complete all the treatments in sequence. The subjects may become fatigued over time. The fatigue factor may tend to systematically impact all the treatment effects of all the subjects, and the researcher cannot control this situation. If the subject is diseased during the first period, regardless of treatment, and the subject is disease free by the time the second period starts, this situation is also a period effect. In many crossover designs, the timing and spacing of periods is relatively loose. In clinical trials, the gap between periods may vary and may depend on when the patient can come in. Large numbers of periods are not suitable for animal feeding and clinical trials with humans. They may be used in psychological experiments with up to 128 periods.

*Sequence effect* is another issue in crossover design. Sequence refers to the order in which the treatments are applied. The possible sets of sequences that might be used in a design depend on the number of the treatments, the length of the sequences, and the aims of the experiment or trial. For instance, with $t$ treatments there are $t!$ possible sequences. The measurement of sequence effect is the average response over the sequences. The simplest design is the two-treatment–two-period, or $2 \times 2$, design. Different treatment sequences are used to eliminate sequence effects. The crossover design cannot accommodate a separate comparison group. Because each experimental unit receives all treatments, the covariates are balanced. The goal of the crossover design is to compare the effects of individual treatments, not the sequences themselves.

## Variance Balance and Unbalance

In the feedstuffs example above, each treatment occurs one time in each experimental unit (or subject) and each of the six three-treatment sequences occurs two times, which confers the property of balance known as *variance balance*. In general, a crossover design is balanced for carryover effect if all possible sequences are used an equal number of times in the experiment, and each treatment occurs equal times in each period and occurs once with each experimental unit. However, this is not always possible; deaths and dropouts may occur in a trial, which can lead to unequal numbers in sequences.

In the example above and Table 1, $A \rightarrow B$ occurs twice, once each in Sequences 1 and 3; and $A \rightarrow C$ occurs once each in Sequences 4 and 5. Similarly, $B \rightarrow A$, $B \rightarrow C$, $C \rightarrow A$, and $C \rightarrow B$ each occur twice. For Experimental Units 1 and 2, $A \rightarrow B$ brings the first-order carryover effect in this experiment and changes the response in the first period following the application of the treatment; similarly, the second-order carryover effect changes the response in the second period following the application of the treatment. A 21-day rest period is used to wash out the effect of treatment before the next treatment is applied. With the absence of carryover effects, the response measurements reflect only the current treatment effect.

In variance balance crossover design, all treatment contrasts are equally precise; for instance, in the example above, we have

$$\text{var}(\tau_A - \tau_B) = \text{var}(\tau_B - \tau_C) = \text{var}(\tau_A - \tau_C),$$

where var = variance and $\tau$ = a treatment group mean.

The contrasts of the carryover effects are also equally precise. The treatment and carryover

effects are negatively correlated in crossover designs with variance balance.

*Balanced crossover* can avoid confounding the period effect with the treatment effects. For example, one group of experimental units received the $A \rightarrow C$, and another group of units received the sequence $C \rightarrow A$. These two treatments are applied in each period, and comparisons of Treatments A and C are independent of comparisons of periods (e.g., Period 1 and Period 2).

The *variance-balanced designs* developed independently by H. D. Patterson and H. L. Lucas, M. H. Quenouille, and I. I. Berenblut contain a large number of periods. These researchers showed that treatment and carryover effects are orthogonal. *Balaam designs* are also balanced, using $t$ treatments, $t^2$ sequences, and only two periods with all treatment combinations, and they are more efficient than the two-period designs of Patterson and Lucas. Adding an extra period or having a baseline observation of the same response variable on each subject can improve the Balaam design. Generally, adding an extra period is better than having baselines. However, the cost is an important factor to be considered by the researcher.

*Variance-unbalanced designs* lack the property that all contrasts among the treatment effect and all contrasts among the carryover effects are of equal precision. For some purposes, the designer would like to have approximate variance balance, which can be fulfilled by the cyclic designs of A. W. Davis and W. B. Hall and the partially balanced incomplete block designs that exist in many of more than 60 designs included in the 1962 paper by Patterson and Lucas. One advantage of the Davis–Hall design is that it can be used for any number of treatments and periods. The efficiencies of Patterson and Lucas and Davis–Hall are comparable, and the designs of Davis and Hall tend to require fewer subjects. *Tied-double-changeover designs* are another type of variance-unbalanced crossover design proposed by W. T. Federer and G. F. Atkinson. The goal of this type is to create a situation in which the variances of differences among treatments are nearly equal to the variances of differences among carryover effects. If Federer and Atkinson's design has $t$ treatments, $p$ periods, and $c$ subjects, and we define two positive integers $q$ and $s$, then $q = (p - 1)/t$ and $s = c/t$. Some combinations of $s$ and $q$ make the design variance balanced.

## Analysis of Different Types of Data

### *Approaches for Normally Distributed Data*

The dependent variables follow a normal distribution. In general, let the crossover design have $t$ treatments and $n$ treatment sequence groups, let $r_i$ = subjects in the $i$th treatment sequence group, and let each group receive treatments in a different order for $p$ treatment periods. Then $y_{ijk}$ is the response value of the $j$th subject of the $i$th treatment sequence in the $k$th period and can be expressed as

$$y_{ijk} = \mu + \alpha_i + \beta_{ij} + \gamma_k + \tau_{d(i,k)} + \lambda_{c(i,k-1)} + \varepsilon_{ijk}$$
$$i = 1, 2, ..n; j = 1, 2, ..r_i; k = 1, 2, ..p; \mathrm{d}, \mathrm{c} = 1, 2, ..t;$$

where $\mu$ is the grand mean, $\alpha_i$ is the effect of the $i$th treatment sequence group, $\beta_{ij}$ is the random effect of the $j$th subject in the $i$th treatment sequence group with variance $\sigma_b^2$, $\gamma_k$ is the $k$th period effect, $\tau_{d(i,k)}$is the direct effect of treatment in period $k$ of sequence group $i$, and $\lambda_{c(i,k-1)}$ is the carryover effect of the treatment in period $k - 1$ of sequence group $i$. Note that $\lambda_{c(i,0)} = 0$ because there is no carryover effect in the first period. And $\beta_{ijk}$ is the random error for the $j$th subject in the $i$th treatment sequence of period $k$ with variance $\sigma^2$. This model is called a mixed model because it contains a random component. In order to simplify, let us denote Treatments A, B, and C as Treatments 1, 2, and 3, respectively. The first observed values in the first and second sequences of the example can be simply written

$$\mathrm{Sequence1}(A \rightarrow B) y_{111} = \mu + \alpha_1 + \beta_{11} + \gamma_1 + \tau_1 + \varepsilon_{111}$$
$$y_{112} = \mu + \alpha_1 + \beta_{11} + \gamma_2 + \tau_2 + \lambda_1 + \varepsilon_{111}$$
$$y_{113} = \mu + \alpha_1 + \beta_{11} + \gamma_3 + \tau_3 + \lambda_2 + \varepsilon_{113}$$

$$\mathrm{Sequence2}(B \rightarrow C) y_{211} = \mu + \alpha_2 + \beta_{21} + \gamma_1 + \tau_2 + \varepsilon_{211}$$
$$y_{212} = \mu + \alpha_2 + \beta_{21} + \gamma_2 + \tau_3 + \lambda_2 + \varepsilon_{212}$$
$$y_{213} \mu + \alpha_2 + \beta_{21} + \gamma_3 + \tau_1 + \lambda_3 + \varepsilon_{213}.$$

From above, it may be seen that there is no carryover effect in the first period, and because only first-order carryovers are considered here, the first-order carryover effects of Treatments A and B are

$\lambda_1$ and $\lambda_2$, respectively, in Sequence 1. Likewise, the first-order carryover effects of Treatments B and C are $\lambda_2$ and $\lambda_3$, respectively, in Sequence 2. Actually, crossover designs are specific repeated measures designs with observed values on each experimental unit that are repeated under different treatment conditions at different time points. The design provides a multivariate observation for each experimental unit.

The univariate analysis of variance can be used for the crossover designs if any of the assumptions of independence, compound symmetry, or the Huynh–Feldt condition are appropriate for the experimental errors. Where independence and compound symmetry are sufficient conditions to justify ordinary least squares, the Huynh–Feldt condition (Type H structure) is both a sufficient and necessary condition for use of ordinary least squares. There are two cases of ANOVA for crossover design: The first is the analysis of variance without carryover effect if the crossover design is a balanced row–column design. The experimental units and periods are the rows and columns of the design, and the direct treatment effects are orthogonal to the columns. The analysis approach is the same as that of a latin square experiment. The second is the analysis of variance with carryover effect, which is treated as a repeated measures split-plot design with the subjects as whole plots and the repeated measures over the $p$ periods as the subplots. The total sum of squares is calculated from between and within subjects' parts. The significance of the carryover effects must be determined before the inference is made on the comparison of the direct effects of treatments.

For normal data without missing values, a least squares analysis to obtain treatment, period, and subject effect is efficient. Whenever there are missing data, within-subject treatment comparisons are not available for every subject. Therefore, additional between-subject information must be used.

In crossover design, the dependent variable may not be normally distributed, such as an ordinal or a categorical variable. The crossover analysis becomes much more difficult if a period effect exists. Two basic approaches can be considered in this case.

### *Approaches of Nonnormal Data*

If the dependent variables are continuous but not normal, the Wilcoxon's rank sum test can be applied to compare the treatment effects and period effects. The crossover should not be used if there is a carryover effect, which cannot be separated from treatment effect or period effects. For this case, bootstrapping, permutation, or randomization tests provide alternatives to normal theory analyses. Gail Tudor and Gary G. Koch described the nonparametric method and its limitations for statistical models with baseline measurements and carryover effects. Actually, this method is an extension of the Mann–Whitney, Wilcoxon's, or Quade's statistics. Other analytical methods are available if there is no carryover effect. For most designs, nonparametric analysis is much more limited than a parametric analysis.

### *Approaches for Ordinal and Binary Data*

New statistical techniques have been developed to deal with longitudinal data of this type over recent decades. These new techniques can also be applied in the analysis of crossover design. A marginal approach using a weighted least square method is proposed by J. R. Landis and others. A generalized estimating equation approach was developed by Kung-Yee Liang and colleagues in 1986. For the binary data, subject effect models and marginal effect models can be applied. Bootstrap and permutation tests are also useful for this type of data. Variance and covariance structures need not necessarily be considered. The different estimates of the treatment effect may be obtained on the basis of different assumptions of the models. Although researchers have developed different approaches for different scenarios, each approach has its own problems or limitations. For example, marginal models can be used to deal with missing data, such as dropout values, but the designs lose efficiency, and their estimates are calculated with less precision. The conditional approach loses information about patient behavior and it is restricted to the logit link function. M. G. Kenward and B. Jones in 1994 gave a more comprehensive discussion of different approaches. Usually, large sample sizes are needed when the data are ordinal or dichotomous, which is a limitation of crossover designs.

Many researchers have tried to find an "optimal" method. J. Kiefer in the 1970s proposed the concept of *universal optimality* and named D-, A-,

and E-optimality. The design is universally optimal and satisfies other optimal conditions, but it is extremely complicated to find a universally optimal crossover design for any given scenario.

*Ying Liu*

*See also* Block Design; Repeated Measures Design

**Further Readings**

Davis, A. W., & Hall, W. B. (1969). Cyclic change-over designs. *Biometrika, 56,* 283–293.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159–174.

Lasserre, V. (1991). Determination of optimal design using linear models in crossover trials. *Statistics in Medicine, 10,* 909–924.

Liang, K. Y., & Zegar, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73,* 13–22.

## Cross-Sectional Design

The methods used to study development are as varied as the theoretical viewpoints on the process itself. In fact, often (but surely not always) the researcher's theoretical viewpoint determines the method used, and the method used usually reflects the question of interest. Age correlates with all developmental changes but poorly explains them. Nonetheless, it is often a primary variable of concern in developmental studies. Hence, the two traditional research designs, longitudinal methods, which examine one group of people (such as people born in a given year), following and reexamining them at several points in time (such as in 2000, 2005, and 2010), and cross-sectional designs, which examine more than one group of people (of different ages) at one point in time. For example, a study of depression might examine adults of varying ages (say 40, 50, and 60 years old) in 2009.

Cross-sectional studies are relatively inexpensive and quick to conduct (researchers can test many people of different ages at the same time), and they are the best way to study age *differences* (not age *changes*). On the other hand, a cross-sectional study cannot provide a very rich picture of development; by definition, such a study examines one small group of individuals at only one point in time. Finally, it is difficult to compare groups with one another, because unlike a longitudinal design, participants do not act as their own controls. Cross-sectional studies are quick and relatively simple, but they do not provide much information about the ways individuals change over time.

As with longitudinal designs, cross-sectional designs result in another problem: the confounding of age with another variable—the cohort (usually thought of as year of birth). *Confounding* is the term used to describe a lack of clarity about whether one or another variable is responsible for observed results. In this case, we cannot tell whether the obtained results are due to age (reflecting changes in development) or some other variable.

Confounding refers to a situation in which the effects of two or more variables on some outcome cannot be separated. Cross-sectional studies confound the time of measurement (year of testing) and age. For example, suppose you are studying the effects of an early intervention program on later social skills. If you use a new testing tool that is very sensitive to the effects of early experience, you might find considerable differences among differently aged groups, but you will not know whether the differences are attributable to the year of birth (when some cultural influence might have been active) or to age. These two variables are confounded.

What can be done about the problem of confounding age with other variables? K. Warner Schaie first identified cohort and time of testing as factors that can help explain developmental outcomes, and he also devised methodological tools to account for and help separate the effects of age, time of testing, and cohort. According to Schaie, age differences among groups represent maturational factors, differences caused by when a group was tested (time of testing) represent environmental effects, and cohort differences represent environmental or hereditary effects or an interaction between the two. For example, Paul B. Baltes and John R. Nesselroade found that differences in the performance of adolescents of the same age on a set of personality tests were related to the year in

which the adolescents were born (cohort) as well as when these characteristics were measured (time of testing).

Sequential development designs help to overcome the shortcomings of both cross-sectional and longitudinal developmental designs, and Schaie proposed two alternative models for developmental research—the *longitudinal sequential design* and the *cross-sectional sequential design*—that avoid the confounding that results when age and other variables compete for attention. Cross-sectional sequential designs are similar to longitudinal sequential designs except that they do not repeat observations on the same people from the cohort; rather, different groups are examined from one testing time to the next. For example participants tested in 2000, 2005, and 2010 would all come from different sets of participants born in 1965. Both of these designs allow researchers to keep certain variables (such as time of testing or cohort) constant while they test the effects of others.

*Neil J. Salkind*

***See also*** Control Variables; Crossover Design; Independent Variable; Longitudinal Design; Research Hypothesis; Research Question; Sequential Design

**Further Readings**

Schaie, K. W. (1992). The impact of methodological changes in gerontology. *International Journal of Human Development & Aging, 35,* 19–29.

Birren, J. E., & Schaie, K. W. (Eds.). (2006). *Handbook of the psychology of aging* (6th ed.). San Diego, CA: Elsevier.

Sneve, M., & Jorde, R. (2008). Cross-sectional study on the relationship between body mass index and smoking, and longitudinal changes in body mass index in relation to change in smoking status: The Tromsø study. (2008). *Scandinavian Journal of Public Health, 36*(4), 397–407.

# Cross-Validation

Cross-validation is a data-dependent method for estimating the prediction error of a fitted model or a trained algorithm. The basic idea is to divide the available data into two parts, called *training data* and *testing data,* respectively. The training data are used for fitting the model or training the algorithm, while the testing data are used for validating the performance of the fitted model or the trained algorithm on predication purpose.

A typical proportion of the training data might be roughly 1/2 or 1/3 when the data size is large enough. The division of the data into training part and testing part can be done naturally or randomly. In some applications, a large enough subgroup of the available data is collected independently of the other parts of the data by different people or institutes, or through different procedures but for similar purposes. Naturally, that part of the data can be extracted and used for testing purpose only. If such a subgroup does not exist, one can randomly draw a predetermined proportion of data for training purposes and leave the rest for testing.

## *K*-Fold Cross-Validation

In many applications, the amount of available data is not large enough for a simple cross-validation. Instead, *K-fold cross-validation* is commonly used to extract more information from the data. Unlike the simple training-and-testing division, the available data are randomly divided into $K$ roughly equal parts. Each part is chosen in turn for testing purposes, and each time the remaining $(K-1)$ parts are used for training purposes. The prediction errors from all the $K$ validations are collected, and the sum is used for cross-validation purposes.

In order to formalize the $K$-fold cross-validation using common statistical notations, suppose the available data set consists of $N$ observations or data points. The $i$th observation includes predictor(s) $x_i$ in scalar (or vector) form and response $y_i$, also known as input and output, respectively. Suppose a random partition of the data divides the original index set $\{1, 2, \ldots, N\}$ into $K$ subsets $I(1), I(2), \ldots, I(K)$ with roughly equal sizes. For the $k$th subset, let $\hat{f}^k_{-}(\cdot)$ be the fitted prediction function based on the rest of the data after removing the $k$th subset. Then the $K$-fold cross-validation targets the average prediction error defined as

$$CV = \frac{1}{N}\sum_{k=1}^{K}\sum_{i \in I(k)} L\left(y_i, \hat{f}^{-k}(\mathrm{x}_i)\right).$$

In the above expression, $CV$ is average prediction error, $L(\cdot, \cdot)$ is a predetermined function, known as the *loss function,* which measures the difference between the observed response $y_i$ and the predicted value $\hat{f}^{-k}(\mathrm{x}_i)$. Commonly used loss functions $L(y, \hat{f})$ include the squared loss function $\left(y - \hat{f}\right)^2$; the absolute loss function $\left|y - \hat{f}\right|$; the $0-1$ loss function, which is 0 if $y = \hat{f}$ and 1 otherwise; and the cross-entropy loss function $-2\log\hat{P}(Y = y_i|\mathrm{x}_i)$.

The result $CV$ of $K$-fold cross-validation depends on the value of $K$ used. Theoretically, $K$ can be any integer between 2 and the data size $N$. Typical values of $K$ include 5 and 10. Generally speaking, when $K$ is small, $CV$ tends to overestimate the prediction error because each training part is only a fraction of the full data set. As $K$ gets close to $N$, the expected bias tends to be smaller, while the variance of $CV$ tends to be larger because the training sets become more and more similar to each other. In the meantime, the cross-validation procedure involves more computation because the target model needs to be fitted for $K$ times. As a compromise, 5-fold or 10-fold cross-validation would be suggested.

When $K = N$ is used, the corresponding cross-validation is known as *leave-one-out cross-validation.* It minimizes the expected bias but can be highly variable. For linear models using squared loss function, one may use the *generalized cross-validation* to overcome the intensive computation problem. Basically, it provides an approximated $CV$ defined as

$$GCV = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{y_i - \hat{f}(\mathrm{x}_i)}{1 - \mathrm{trace(S)}/N}\right]^2.$$

Here S is an $N \times N$ matrix setting the fitting equation $\hat{y} = \mathrm{S}y$, and trace(S) is the sum of the diagonal elements of S.

## Cross-Validation Applications

The idea of cross-validation can be traced back to the 1930s. It was further developed and refined in the 1960s. Nowadays, it is widely used, especially when the data are unstructured or fewer model assumptions could be made. The cross-validation procedure does not require distribution assumptions, which makes it flexible and robust.

### *Choosing Parameter*

One of the most successful applications of cross-validation is to choose a smoothing parameter or penalty coefficient. For example, a researcher wants to find the best function $f$ for prediction purposes but is reluctant to add many restrictions. To avoid the overfitting problem, the researcher tries to minimize a penalized residual sum of squares defined as follows:

$$RSS(f, \lambda) = \sum_{i=1}^{N}[y_i - f(x_i)]^2 + \lambda\int [f''(x)]^2\, dx.$$

Then the researcher needs to determine which $\lambda$ should be used, because the solution changes along with it. For each candidate $\lambda$, the corresponding fitted $f$ is defined as the solution minimizing $RSS(f, \lambda)$. Then the cross-validation can be applied to estimate the prediction error $CV(\lambda)$. The optimal $\lambda$ according to cross-validation is the one that minimizes $CV(\lambda)$.

### *Model Selection*

Another popular application of cross-validation is to select the best model from a candidate set. When multiple models, or methods, or algorithms are applied to the same data set, a frequently asked question is which one is the best. Cross-validation provides a convenient criterion to evaluate their performance. One can always divide the original data set into training part and testing part. Fit each model based on training data and estimate the prediction error of the fitted model based on the testing data. Then the model that attains the minimal prediction error is the winner.

Two major concerns need to be addressed for the procedure. One of them is that the estimated prediction error may strongly depend on how the data are divided. The value of *CV* itself is random if it is calculated on the basis of a random partition. In order to compare two random *CV*s, one may need to repeat the whole procedure for many times and compare the *CV* values on average. Another concern is that the selected model may succeed for one data set but fail for another because cross-validation is a data-driven method. In practice, people tend to test the selected model again when additional data are available via other sources before they accept it as the best one.

Cross-validation may also be needed during fitting a model as part of the model selection procedure. In the previous example, cross-validation is used for choosing the most appropriate smoothing parameter $\lambda$. In this case, the data may be divided into three parts: training data, validation data, and testing data. One may use the validation part to choose the best $\lambda$ for fitting the model and use the testing part for model selection.

*Jie Yang*

***See also*** Bootstrapping; Jackknife

**Further Readings**

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall/CRC.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York: Springer.

# Cumulative Frequency Distribution

Cumulative frequency distributions report the frequency, proportion, or percentage of cases at a particular score or less. Thus, the cumulative frequency of a score is calculated as the frequency of occurrence of that score plus the sum of the frequencies of all scores with a lower value. Cumulative frequency distributions are usually displayed with the aid of tables and graphs and may be put together for both ungrouped and grouped scores.

## Cumulative Frequency Tables for Distributions With Ungrouped Scores

A cumulative frequency table for distributions with ungrouped scores typically includes the scores a variable takes in a particular sample, their frequencies, and the cumulative frequency. In addition, the table may include the cumulative relative frequency or proportion, and the cumulative percentage frequency. Table 1 illustrates the frequency, cumulative frequency, cumulative relative frequency, and cumulative percentage frequency for a set of data showing the number of credits a sample of students at a college have registered for in the autumn quarter.

The cumulative frequency is obtained by adding the frequency of each observation to the sum of the frequencies of all previous observations (which is, actually, the cumulative frequency on the previous row). For example, the cumulative frequency for the first row in Table 1 is 1 because there are no previous observations. The cumulative frequency for the second row is $1 + 0 = 1$. The cumulative frequency for the third row is $1 + 2 = 3$. The cumulative frequency for the fourth row is $3 + 1 = 4$, and so on. This means that four students have registered for 13 credits or fewer in the autumn quarter. The cumulative frequency for the last observation must equal the number of observations included in the sample.

Cumulative relative frequencies or cumulative proportions are obtained by dividing each cumulative frequency by the number of observations. Cumulative proportions show the proportion of observations that fulfill a particular criterion or less. For example, the proportion of students who have registered for 14 credits or fewer in the autumn quarter is 0.60. The cumulative proportion for the last observation (last row) is always 1.

Cumulative percentages are obtained by multiplying the cumulative proportions by 100. Cumulative percentages show the percentage of observations that fulfill a certain criterion or less. For example, 40% of students have registered for 13 credits or fewer in the autumn quarter. The cumulative percentage of the last observation (the last row) is always 100.

**Table 1** Cumulative Frequency Distribution of the Number of Credits Students Have Registered for in the Autumn Quarter

| *Number of credits* $y$ | *Frequency* $f$ | *Cumulative frequency* $cum(f)$ | *Cumulative relative frequency* $cumr(f) = cum(f)/n$ | *Cumulative percentage frequency* $cump(f) = 100 * cumr(f)$ |
|---|---|---|---|---|
| 10 | 1 | 1 | 0.10 | 10.00 |
| 11 | 0 | 1 | 0.10 | 10.00 |
| 12 | 2 | 3 | 0.30 | 30.00 |
| 13 | 1 | 4 | 0.40 | 40.00 |
| 14 | 2 | 6 | 0.60 | 60.00 |
| 15 | 4 | 10 | 1.00 | 100.00% |
| | $n = 10$ | | | |

**Table 2** Grouped Cumulative Frequency Distribution of Workers' Salaries

| *Workers' salaries* $y$ | *Frequency* $f$ | *Cumulative frequency* $cum(f)$ | *Cumulative relative frequency* $cumr(f) = cum(f)/n$ | *Cumulative percentage frequency* $cump(f) = 100 * cumr(f)$ |
|---|---|---|---|---|
| \$15,000–\$19,999 | 4 | 4 | 0.27 | 26.67 |
| \$20,000–\$24,999 | 6 | 10 | 0.67 | 66.67 |
| \$25,000–\$29,999 | 3 | 13 | 0.87 | 86.67 |
| \$30,000–\$34,999 | 0 | 13 | 0.87 | 86.67 |
| \$35,000–\$39,999 | 2 | 15 | 1.00 | 100.00% |
| | $n = 15$ | | | |

## Grouped Cumulative Frequency Distributions

For distributions with grouped scores, the cumulative frequency corresponding to each class equals the frequency of occurrence of scores in that particular class plus the sum of the frequencies of scores in all lower classes (which is, again, the cumulative frequency on the previous row). Grouped cumulative frequency distributions are calculated for continuous variables or for discrete variables that take too many values for the list of all possible values to be useful. The cumulative distribution table for this case is very similar to Table 1, and the entries in the table are computed in a similar way. The only difference is that instead of individual scores, the first column contains classes. Table 2 is an example of a cumulative frequency table for a sample of workers' salaries at a factory.

The cumulative frequency of a particular class is calculated as the frequency of the scores in that class plus the sum of the frequencies of scores in all lower classes. Cumulative frequencies in Table 2 show the number of workers who earn a certain amount or money or less. For example, the reader learns that 10 workers earn \$24,999 or less.

Cumulative relative frequencies or proportions show the proportion of workers earning a certain amount of money or less and are calculated by dividing the cumulative frequency by the number of observations. For example, the proportion of workers earning \$29,999 or less is 0.87.

Finally, cumulative percentages show the percentage of workers earning a certain amount of money or less, and these percentages are obtained by multiplying the cumulative relative frequency by 100. For example, 86.67% of workers earn \$34,999 or less. The last row shows that 15 workers, representing all the workers included in the analyzed sample (100%), earn \$39,999 or less. The cumulative proportion of the workers in this last class is, logically, 1.

## Graphing Cumulative Distributions of Discrete Variables

Cumulative frequencies, cumulative proportions, and cumulative percentages of a discrete variable can all be represented graphically. An upper right quadrant of a two-dimensional space is

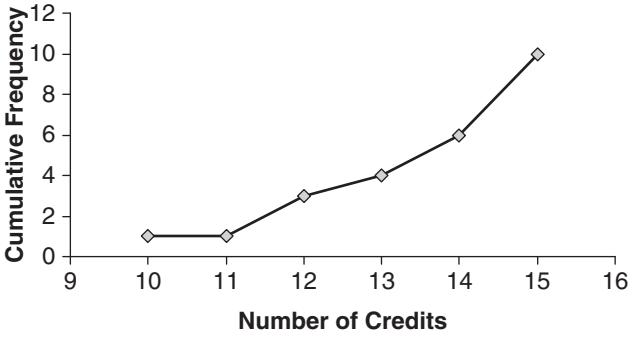



**Figure 1** Cumulative Frequency Distribution of the Number of Credits Students Have Registered for in the Autumn Quarter

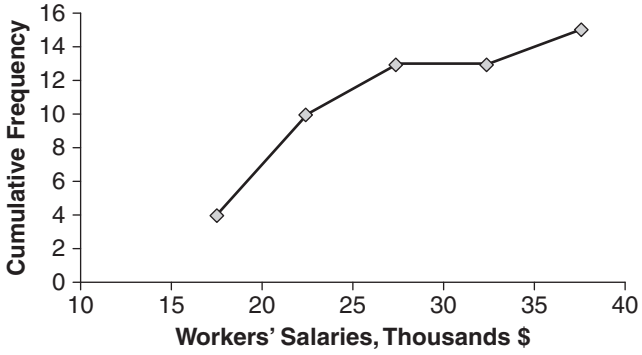



**Figure 2** Grouped Cumulative Frequency Distribution of Workers' Salaries

typically used to display a cumulative distribution. The quadrant is bounded by an *x*-axis and a *y*-axis. The *x*-axis is positioned horizontally and usually corresponds to the scores that the variable can take. The *y*-axis indicates the cumulative frequency, proportion, or percentage. Cumulative distributions may be graphed using histograms or, more commonly, polygons.

To exemplify, the number of credits from Table 1 is plotted in Figure 1. The number of credits the students have registered for is represented on the *x*-axis. The cumulative frequency is represented on the *y*-axis. Using each score and its corresponding cumulative frequency, a number of points are marked inside the quadrant. They are subsequently joined to form a polygon that represents the cumulative frequency distribution graph.

## Graphing Grouped Cumulative Distributions

Grouped cumulative frequency distributions may also be graphed via both histograms and polygons. Again, it is more common to use cumulative frequency (or proportion or percentage) polygons. To draw the polygon for a grouped distribution, the points represented by the midpoint of each class interval and their corresponding cumulative frequency are connected.

To exemplify, the cumulative frequency distribution described in Table 2 is graphically illustrated in Figure 2.

## Shape of Cumulative Distributions

Cumulative polygons are often referred to as ogives. They climb steeply in regions where classes include many observations, such as the middle of a bell-shaped distribution, and climb slowly in regions where classes contain fewer observations. Consequently, frequency distributions with bell-shaped histograms generate S-shaped cumulative frequency curves. If the distribution is bell shaped, the cumulative proportion equals 0.50 at the average of the *x* value, which is situated close to the midpoint of the horizontal range. For positive skews, the cumulative proportion reaches 0.50 before the *x* average. On the contrary, for negative skews, the cumulative proportion reaches 0.50 further away to the right, after the *x* average.

Many statistical techniques work best when variables have bell-shaped distributions. It is, however, essential to examine the actual form of the data distributions before one uses these techniques. Graphs provide the simplest way to do so, and cumulative polygons often offer sufficient information.

## Using Cumulative Distributions in Practice

Cumulative frequency (or proportion or percentage) distributions have proved useful in a multitude of research spheres. Typically, cumulative percentage curves allow us to answer questions such as, What is the middle score in the distribution? What is the dividing point for the top 30% of the group? Consequently, cumulative percentage curves are

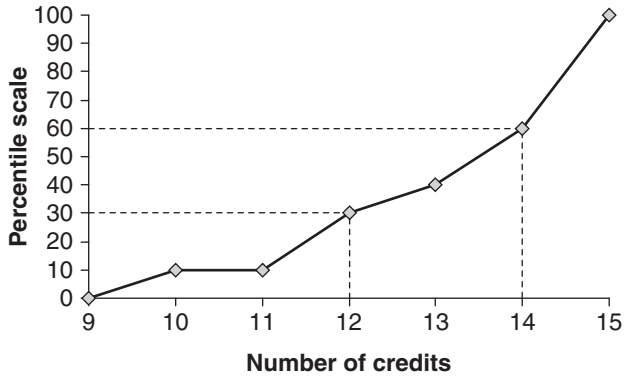



**Figure 3** Cumulative Percentage Frequency Distribution of the Number of Credits Students Have Registered for in the Autumn Quarter

used by instructors in schools, universities, and other educational centers to determine the proportion of students who scored less than a specified limit. Doctors' offices may use cumulative weight and height distributions to investigate the percentage of children of a certain age whose weight and height are lower than a certain standard. All this is usually done by using cumulative frequency distributions to identify percentiles and percentile ranks. Consequently, the cumulative frequency curve is also known as the *percentile curve*.

## Finding Percentiles and Percentile Ranks From Cumulative Distributions

A percentile is the score at or below which a specified percentage of scores in a distribution falls. For example, if the 40th percentile of an examination is 120, it means that 40% of the scores on the examination are equal to or less than 120. The percentile rank of a score indicates the percentage of scores in the distribution that are equal to or less than that score. Referring to the above example, if the percentile rank of a score of 120 is 40, it means that 40% of the scores are equal to or less than 120.

Percentiles and percentile ranks may be determined with certain formulas. Nevertheless, as already mentioned, the cumulative distribution polygon is often used to determine percentiles and percentile ranks graphically. To exemplify, Figure 3 shows the cumulative percentage frequency distribution for the data in Table 1, with the cumulative percentage/percentile scale represented on the *y*-axis and the number of credits on the *x*-axis.

To determine the percentile rank corresponding to 12 credits, a perpendicular is erected from the *x*-axis at point 12 until it meets the cumulative polygon. A horizontal line is then drawn from this point until it meets the *y*-axis, which is at point 30 on the percentile scale. This means that 30% of the students have registered for 12 credits or fewer for the autumn quarter. The cumulative percentage frequency data shown in Table 1 confirm this result. The method may be employed the other way around as well. The percentile rank may be used to determine the corresponding percentile. Drawing a horizontal line from the point 60 on the percentile scale to the cumulative polygon and then a perpendicular on the *x*-axis from the point of intersection, the point (0, 14) is met. In other words, the 60th percentile corresponds to 14 credits, showing that 60% of the students have registered for 14 credits or fewer. The midpoint in a distribution is often known as the 50th percentile, the median, or the second quartile. It represents the point below which we find 50% of the scores and above which the other 50% of the scores are located. The 10th percentile is called the first decile, and each multiple of 10 is referred to as a decile. The 25th percentile is the first quartile, and the 75th percentile is the third quartile. This is a very simple method to determine the percentile rank corresponding to a particular score. However, the method may not be always accurate, especially when a very fine grid is not available.

*Oana Pusa Mihaescu*

***See also*** Descriptive Statistics; Distribution; Frequency Distribution; Frequency Table; Histogram; Percentile Rank

### Further Readings

Downie, N. M., & Heath, R. W. (1974). *Basic statistical methods*. New York: Harper & Row.

Fielding, J. L., & Gilbert, G. N. (2000). *Understanding social statistics*. Thousand Oaks, CA: Sage.

Fried, R. (1969). *Introduction to statistics*. New York: Oxford University Press.

Hamilton, L. (1996). *Data analysis for social sciences: A first course in applied statistics.* Belmont, CA: Wadsworth.

Kiess, H. O. (2002). *Statistical concepts for the behavioral sciences.* Boston: Allyn and Bacon.

Kolstoe, R. H. (1973). *Introduction to statistics for the behavioral sciences*. Homewood, IL: Dorsey Press.

Lindquist, E. F. (1942). *A first course in statistics: Their use and interpretation in education and psychology.* Cambridge, MA: Riverside Press.

# D

## Databases

One of the most efficient and increasingly common methods of investigating phenomena in the education and social sciences is the use of databases. Large-scale databases generally comprise information collected as part of a research project. Information included in databases ranges from survey data from clinical trials to psychoeducational data from early childhood projects. Research projects from which databases are derived can be longitudinal or cross-sectional in nature, use multiple or individual informants, be nationally representative or specific to a state or community, and be primary data for the original researcher or secondary data for individuals conducting analysis at a later time. This entry explores the benefits and limitations of using databases in research, describes how to locate databases, and discusses the types of databases and the future of the use of databases in research.

### Benefits

The primary advantage of using databases for research purposes is related to economics. Specifically, since databases consist of information that has already been collected, they save researchers time and money because the data are readily available. As with many investigators, the primary hindrance to conducting original field research is limited monetary resources. Collecting data from large samples is time-consuming, and many direct and indirect costs are associated with obtaining access to specific populations for collection of specific data. This limitation is eliminated by using large-scale databases. Depending on the topic of interest, the use of databases provides researchers access to randomly sampled and nationally representative populations.

Databases also provide researchers with access to populations they may not have had access to individually. Specifically, the recruitment of individuals from diverse backgrounds (e.g., Black, Latino) has generally been a problem in the social and medical sciences due to historical issues centering on mistrust of researchers (e.g., the Tuskegee Experiment). While this is the case, databases such as the National Institute of Mental Health–funded Collaborative Psychiatric Epidemiology Surveys (CPES) provide access to diverse subjects. Specifically, CPES joins together three nationally representative surveys: the National Comorbidity Survey Replication (NCS-R), the National Survey of American Life (NSAL), and the National Latino and Asian American Study (NLAAS). These studies collectively provide the first national data with sufficient power to investigate cultural and ethnic influences on mental disorders. Although existing databases offer numerous benefits, they have limitations as well.

### Limitations

The key limitation of using databases is that questions and the theoretical orientation of the original

researchers may not be congruent with those of the secondary investigator. So if a researcher was not part of the original research team, the conceptualization of the constructs of interest in the database may not be to his or her liking. Although numerous available databases encompass a variety of topics, this limitation can be virtually impossible to ignore. To combat it, researchers generally undertake the task of recoding questions and variables to fit their research questions of interest.

In addition to question conceptualization problems, another limitation of databases is the date the data were collected. Specifically, if an individual uses a database that is dated, this may impact his or her ability to generalize his or her findings to the present day. This threat to internal validity can be lessened if researchers use the most up-to-date database on their topic of interest. An example of this is the U.S. Department of Education–funded Education Longitudinal Study of 2002. This study is a direct follow-up to the National Education Longitudinal Study of 1988. Although the 1988 study resulted in a high-quality, longitudinal database with significant policy implications, stakeholders realized that the database was dated, and the result was the initiation of the 2002 study.

## How to Locate Databases

Due to the quantity of information and the manpower necessary to conduct projects of this scope, the funding of large-scale research generally comes from governmental entities such as the National Institutes of Health (NIH) and the U.S. Department of Education. Since these institutions are taxpayer funded, these large-scale databases are generally available for free to researchers. However, because of the sensitive and personally identifiable information that can be ascertained from these databases, researchers must obtain a restricted-use data license before many databases can be accessed. A restricted-use license consists of a justification of the need for the restricted-use data, an agreement to keep the data safe from unauthorized disclosures at all times, and an agreement to participate fully in unannounced, unscheduled inspections by the U.S. Department of Education or NIH security officials to ensure compliance with the terms of the license and the security procedures and plan.

## Types of Databases

### *Early Childhood Databases*

#### Early Childhood Longitudinal Study

The Early Childhood Longitudinal Study (ECLS) consists of two overlapping cohorts: a birth cohort and a kindergarten cohort, better known as the ECLS-B and the ECLS-K, respectively. These databases include children who were followed from birth through kindergarten entry and from kindergarten through the eighth grade, respectively. The nationally representative ECLS-B consists of 14,000 children born in the year 2001. The children participating in the study come from diverse socioeconomic and ethnic backgrounds, with oversamples of Asian and Pacific Islander children, American Indian and Alaska Native children, Chinese children, twins, and low and very low birth weight children. Information about these children was collected when the children were approximately 9 months old, 2 years old (2003), and in preschool (1 year away from kindergarten, fall 2005). In fall 2006, data were collected from all participating sample children, 75% of whom were expected to be age eligible for kindergarten. In fall 2007, data were collected from the remaining 25% of participating sample children, who were newly eligible for kindergarten.

The ECLS-K is a nationally representative sample of kindergartners, their families, their teachers, and their schools all across the United States. Information was collected in the fall and spring of kindergarten (1998–1999); the fall and spring of first grade (1999–2000); and the spring of third grade (2002), fifth grade (2004), and eighth (2007) grade. It was designed to provide comprehensive and reliable data that may be used to describe and better understand children's development and experiences in the elementary and middle school grades and how their early experiences relate to their later development, learning, and experiences in school. The multifaceted data collected across the years allow researchers and policy makers to study how various student, home, classroom, school, and community factors at various points in a child's life relate to the child's cognitive and social development.

#### Head Start Family and Child Experiences Survey

The Head Start Family and Child Experiences Survey (FACES) provides longitudinal data on

the characteristics, experiences, and outcomes of Head Start children and families, as well as the characteristics of the Head Start programs that serve them. In 1997 the Department of Health and Human Services, Administration for Children and Families, commissioned FACES. The success of the original FACES database prompted follow-ups, and currently there are four FACES databases: FACES 1997, 2000, 2003, and 2006. Each cohort has included a nationally representative sample of Head Start children and their families. FACES has several major objectives which include studying the relationship among family, preschool, and school experiences; children's academic development in elementary school; and the developmental progression of children as they progress from Head Start to elementary school.

#### NICHD Study of Early Child Care and Youth Development

The National Institute of Child Health and Human Development (NICHD) Study of Early Child Care and Youth Development (SECCYD) is a comprehensive longitudinal study initiated by the NICHD to answer questions about the relationships between child care experiences, child care characteristics, and children's developmental outcomes. The SECCYD data are from 1,364 families, followed since their infant's birth in 1991. The study covers demographic, family, maternal, paternal, and caregiver characteristics; child social and emotional outcomes; language development; cognitive skills; school readiness; and growth and health measures. The study was conducted in four phases, based on the ages of the children. Phase I of the study was conducted from 1991 to 1994, following the children from birth to age 3 years. Phase II of the study was conducted between 1995 and 2000 to follow the 1,226 children continuing to participate from age 3 through their 2nd year in school. Phase III of the study was conducted between 2000 and 2005 to follow more than 1,100 of the children through their 7th year in school. Phase IV will follow more than 1,000 of the original families through age 15.

### *Adolescent Databases*

#### Education Longitudinal Study

The Education Longitudinal Study (ELS) of 2002 is part of the National Center for Education Statistics' National Education Longitudinal Studies program, which also includes three completed studies: the National Longitudinal Study of the High School Class of 1972, the High School and Beyond longitudinal study of 1980, and the National Education Longitudinal Study of 1988. The ELS database consists of a nationally representative sample of students tracked from the time they were high school sophomores until they enter postsecondary education and the labor market. As such, this database allows researchers to access information about individuals from the time they are adolescents until their mid- to late 20s. Data from this study are derived from information obtained from students, school records, parents, teachers, and high school administrators.

#### National Longitudinal Study of Adolescent Health

The National Longitudinal Study of Adolescent Health (Add Health) is a nationally representative, school-based, longitudinal study that explores variables related to health behaviors for adolescents. The Add Health database provides researchers with information on how social contexts (e.g., families, schools, and neighborhoods) influence adolescents' health and risk behaviors. Add Health was started in 1994 by way of a grant from NICHD. Data at the individual, family, school, and community levels were collected in two waves between 1994 and 1996. In 2001 and 2002, Add Health respondents, 18 to 26 years old, were reinterviewed in a third wave to investigate the influence that adolescence has on young adulthood.

### *Special Population Databases*

#### Pre-Elementary Education Longitudinal Study

The Pre-Elementary Education Longitudinal Study (PEELS) is part of a group of studies on the experiences, special services, and outcomes of children, youth, and young adults with disabilities. The children were 3 to 5 years old at the start of the study. The purpose of PEELS is a better understanding of the answers to the following questions:

What are the characteristics of children receiving preschool special education? What preschool programs and services do they receive? How are transitions between early intervention (programs for children from birth to 3 years old) and preschool, and between preschool and elementary school? How do these children function and perform in preschool, kindergarten, and early elementary school? Which child, school program, and/or special service characteristics are associated with better results in school? Data collection began in fall 2003 and was repeated in winter 2005, 2006, 2007, and 2009.

#### National Longitudinal Transition Study-2

The National Longitudinal Transition Study-2 (NLTS-2) is a 10-year study documenting the characteristics, experiences, and outcomes of a nationally representative sample of more than 11,000 youth who were ages 13 through 16 and were receiving special education services in Grade 7 or above when the study began in 2001. The NLTS-2, funded by the U.S. Department of Education, is a follow-up of the original National Longitudinal Transition Study, conducted from 1985 through 1993. Information in this database is derived from children with disabilities, their parents, and their schools. The study follows these children through young adulthood.

#### Special Education Elementary Longitudinal Study

The Special Education Elementary Longitudinal Study (SEELS) is a study of school-age children who were in special education. SEELS was funded by the U.S. Department of Education and is part of the national assessment of the 1997 Individuals With Disabilities Education Act. SEELS involves a large, nationally representative sample of students in special education who were age 6 through 12 in 1999. Students were selected randomly from rosters of students in special education provided by local education agencies and state-operated special schools for the deaf and blind that agreed to participate in the study. Beginning in the year 2000 and concluding in 2006, SEELS documented the school experiences of a national sample of students as they move from elementary to middle school and from middle to high school.

### Future of Database Research

As mentioned previously there are numerous databases for the study of virtually any phenomenon. While this may be the case, the National Children's Study may be the most ambitious database project. The development of this study was spurred by the 1998 President's Task Force on Environmental Health and Safety Risks to Children recommendation that a large prospective epidemiologic study of U.S. children be done. As such, the U.S. Congress, through the Children's Health Act of 2000, gave the NICHD the task of conducting a national longitudinal study of environmental influences (including physical, chemical, biological, and psychosocial influences) on children's health and development. Funded by the National Institute of Environmental Health Sciences, the Centers for Disease Control and Prevention (CDC), and the U.S. Environmental Protection Agency, the study will cost an estimated $2.3 billion over 30 years and was initiated in 2007. Its purpose is to understand factors related to environmental risks and individual susceptibility factors for asthma, birth defects, dyslexia, attention deficit/hyperactivity disorder, autism, schizophrenia, and obesity, as well as for adverse birth outcomes. Participants for this study are a nationally representative sample of 100,000 children. These children will be followed from conception to 21 years of age and examines environmental exposures. It includes genetic information and chemical analysis from the families' communities. The database is intended to produce an extremely rich set of information for the study of human development.

Historically, investigators have been encouraged to collect their own data for research purposes. While this is still true, the cost associated with conducting independent field research and difficulty gaining access to diverse samples requires individuals to use other methods for access to data. One of these methods is to use large-scale databases. As demonstrated by published research, the potential use of existing databases to answer research questions is virtually unlimited. Given the widespread availability of databases from multiple fields and the ongoing efforts of governmental agencies to fund

these endeavors, the use of databases will remain a staple in research activities.

*Scott Graves*

***See also*** Primary Data Source; Secondary Data Source

**Further Readings**

Anderson, C., Fletcher, P., & Park, J. (2007). *Early Childhood Longitudinal Study, Birth Cohort (ECLS–B) Psychometric Report for the 2-year Data Collection* (NCES 2007–084). Washington, DC: National Center for Education Statistics.

Landrigan, P., Trasande, L., Thorpe, L., Gwynn, C., Lioy, P., D'Alton, M., et al. (2006). The national children's study: A 21-year prospective study of 10000 American children. *Pediatrics, 118,* 2173–2186.

Markowitz, J., Carlson, E., Frey, W., Riley, J., Shimshak, A., Heinzen, H., et al. (2006). *Preschoolers' characteristics, services, and results: Wave 1 overview report from the Pre-Elementary Education Longitudinal Study (PEELS).* Rockville, MD: Westat.

NICHD Early Child Care Research Network. (1993).Child care debate: Transformed or distorted?. *American Psychologist, 48,* 692–693.

Pennell, B., Bowers, A., Carr, D., Chardoul, S., Cheung, G., Dinkelmann, K., et al. (2004). The development and implementation of the National Comorbidity Survey Replication, the National Survey of American Life, and the National Latino and Asian American Survey. *International Journal of Methods in Psychiatric Research, 13,* 241–269.

Rock, D., & Pollack, J. (2002). *Early Childhood Longitudinal Study—Kindergarten Class of 1998–99 (ECLS-K) psychometric report for kindergarten through first grade (NCES 2002–05).* Washington, DC: National Center for Education Statistics.

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W., & Shavelson, R. (2007). *Estimating causal effects: Using experimental and observational designs.* Washington, DC: American Educational Research Association.

U.S. Department of Health and Human Services. (2002). *A descriptive study of Head Start families: FACES technical report I.* Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families.

Wagner, M., Kutash, K., Duchnowski, A., & Epstein, M. (2005). The special education elementary longitudinal study and the national longitudinal transition study: Study designs and implications for children and youth with emotional disturbance. *Journal of Emotional and Behavioral Disorders, 13,* 25–41.

Wagner, M., Marder, C., Levine, P., Cameto, P., Cadwallader, T., & Blackorby, J. (2003). *The individual and household characteristics of youth with disabilities: A report from the National Longitudinal Transition Study-2 (NLTS2).* Menlo Park, CA: SRI International.

# Data Cleaning

Data cleaning, or *data cleansing,* is an important part of the process involved in preparing data for analysis. Data cleaning is a subset of *data preparation,* which also includes scoring tests, matching data files, selecting cases, and other tasks that are required to prepare data for analysis.

Missing and erroneous data can pose a significant problem to the reliability and validity of study outcomes. Many problems can be avoided through careful survey and study design. During the study, watchful monitoring and data cleaning can catch problems while they can still be fixed. At the end of the study, multiple imputation procedures may be used for data that are truly irretrievable.

The opportunities for data cleaning are dependent on the study design and data collection methods. At one extreme is the anonymous Web survey, with limited recourse in the case of errors and missing data. At the other extreme are longitudinal studies with multiple treatment visits and outcome evaluations. Conducting data cleaning during the course of a study allows the research team to obtain otherwise missing data and can prevent costly data cleaning at the end of the study. This entry discusses problems associated with data cleaning and their solutions.

## Types of "Dirty Data"

Two types of problems are encountered in data cleaning: missing data and errors. The latter may be the result of respondent mistakes or data entry errors. The presence of "dirty data" reduces the reliability and validity of the measures. If responses are missing or erroneous, they will not be reliable over time. Because reliability sets the upper bound for validity, unreliable items reduce validity.

**Table 1** Data Errors and Missing Data

| *Variable* | *"True" Data* | *Incomplete, Incorrect, or Missing Data* |
|---|---|---|
| Name | Maria Margaret Smith | Maria Smith |
| Date of birth | 2/19/1981 | 1981 |
| Sex | F | M |
| Ethnicity | Hispanic and Caucasian | |
| Education | B.A., Economics | College |
| Place of birth | Nogales, Sonora, Mexico | Nogales |
| Annual income | $50,000 | |

### *Missing Data*

Missing data reduce the sample size available for the analyses. An investigator's research design may require 100 respondents in order to have sufficient power to test the study hypotheses. Substantial effort may be required to recruit and treat 100 respondents. At the end of the study, if there are 10 important variables, with each variable missing only 5% of the time, the investigator may be reduced to 75 respondents with complete data for the analyses. Missing data effectively reduce the power of the study. Missing data can also introduce bias because questions that may be embarrassing or reveal anything illegal may be left blank. For example, if some respondents do not answer items about income, place of birth (for immigrants without documents), or drug use, the remaining cases with complete data are a biased sample that is no longer representative of the population.

### *Data Errors*

Data errors are also costly to the study because lowered reliability attenuates the results. Respondents may make mistakes, and errors can be introduced during data entry. Data errors are more difficult to detect than missing data. Table 1 shows examples of missing data (ethnicity, income), incomplete data (date and place of birth), and erroneous data (sex).

### *Causes*

All measuring instruments are flawed, regardless of whether they are in the physical or social sciences. Even with the best intentions, everyone makes errors. In the social sciences, most measures are self-report. Potentially embarrassing items can result in biased responses. Lack of motivation is also an important source of error. For example, respondents will be highly motivated in high-stakes testing such as the College Board exams but probably do not bring the same keen interest to one's research study.

## Solutions and Approaches

Data problems can be prevented by careful study design and by pretesting of the entire research protocol. After the forms have been collected, the task of data cleaning begins. The following discussion of data cleaning is for a single paper-and-pencil survey collected in person. Data cleaning for longitudinal studies, institutional data sets, and anonymous surveys is addressed in a later section.

### *Missing Data*

The best approach is to fill in missing data as soon as possible. If a data collection form is skimmed when it is collected, the investigator may be able to ask questions at that time about any missing items. After the data are entered, the files can be examined for remaining missing values. In many studies, the team may be able to contact the respondent or fill in basic data from memory. Even if the study team consists of only the principal investigator, it is much easier to fill in missing data after an interview than to do so a year later. At that point, the missing data may no longer be retrievable.

### *Data Entry Errors*

A number of helpful computer procedures can be used to reduce or detect data entry errors, such as double entry or proactive database design. Double entry refers to entering the data twice, in order to ensure accuracy. Careful database design includes structured data entry screens that are limited to specified formats (dates, numbers, or text) or ranges (e.g., sex can only be M or F, and age must be a number less than 100).

#### Perils of Text Data

Text fields are the default for many database management systems, and a new user may not be aware of the problems that text data can cause at the end of a project. For example, "How many days did you wait before you were seen by a physician?" should require a number. If the database allows text data, the program will accept "10," "ten days," "10 days," "5 or 6," "half a day," "2–3," "about a month," "not sure," "NA," and so on. Before analysis can proceed, every one of these answers must be converted to a number. In large studies, the process of converting text to numbers can be daunting. This problem can be prevented by limiting fields to numeric inputs only.

#### Logically Impossible Responses

Computer routines can be used to examine the data logically and identify a broad range of errors. These validation routines can detect out-of-range values and logical inconsistencies; there are some combinations that are not possible in a data set. For example, all the dates for data collection should fall between the project's starting date and ending date. In a study of entire families, the children should be younger than their parents, and their ages should be somewhat consistent with their reported grade levels. More sophisticated procedures, such as Rasch modeling techniques, can be used to identify invalid surveys submitted by respondents who intentionally meddled with the system (for example, by always endorsing the first response).

### *Logistics*

The amount of time required for data cleaning is dependent on who cleans the data and when the data cleaning takes place. The best people to clean the data are the study team, and the best time is while the study is under way. The research staff is familiar with the protocol and probably knows most of the study participants. A data set that is rife with problems can require a surprising amount of time for data cleaning at the end of a study.

## Longitudinal Studies and Other Designs

Anonymous surveys, secondary data sets, longitudinal studies with multiple forms collected on multiple occasions, and institutional databases present different data-cleaning problems and solutions.

Anonymous surveys provide few opportunities for data cleaning. It is not possible to ask the respondent to complete the missing items. Data cleaning will be limited to logic checks to identify inconsistent responses and, in some cases, using the respondent's answers to other questions.

Secondary data sets, such as those developed and maintained by the Centers for Disease Control and Prevention, will be fully documented, with online files for the data collection forms, research protocols, and essential information about data integrity. The data collection instruments have been designed to minimize errors, and the majority of possible data cleaning will be complete.

Institutional databases can be an excellent source of information for addressing research questions. Hospital billing and claims files and college enrollment records have often been in place for decades. However, these systems were designed to meet an institution's ongoing requirements and were rarely planned with one's research study in mind. *Legacy databases* present some unique challenges for data cleaning.

In contrast to secondary data analyses, institutional databases may have limited documentation available to the outside researcher. Legacy databases usually change in order to meet the needs of the institution. The records of these changes may be unavailable or difficult for nonprogrammers to follow. User-initiated changes may be part of the institutional lore. Data accuracy may be high for critical data fields but lower for other variables, and a few variables may be entirely blank. Variable formats may change across data files or tables. For example, the same item in three data sets may have different field characteristics, with a leading space in the first, a leading zero in the second, and neither in the third. Data-cleaning approaches may involve significant time examining the data. The accuracy of the final product will need to be verified with knowledgeable personnel.

Studies with multiple instruments allow the investigators to reconstruct additional missing and erroneous data by triangulating data from the

other instruments. For example, one form from a testing battery may be missing a date that can be inferred from the rest of the packet if the error is caught during data entry.

*Longitudinal studies* provide a wealth of opportunity to correct errors, provided early attention to data cleaning and data entry have been built into the study. Identification of missing data and errors while the respondent is still enrolled in the study allows investigators to "fill in the blanks" at the next study visit. Longitudinal studies also provide the opportunity to check for consistency across time. For example, if a study collects physical measurements, children should not become shorter over time, and measurements should not move back and forth between feet and meters.

Along with opportunities to catch errors, studies with multiple forms and/or multiple assessment waves also pose problems. The first of these is the file-matching problem. Multiple forms must be matched and merged via computer routines. Different data files or forms may have identification numbers and sorting variables that do not exactly match, and these must be identified and changed before matching is possible.

### *Documentation*

Researchers are well advised to keep a log of data corrections in order to track changes. For example, if a paper data collection form was used, the changes can be recorded on the form, along with date the item was corrected. Keeping a log of data corrections can save the research team from trying to clean the same error more than once.

### *Data Integrity*

Researcher inexperience and the ubiquitous lack of resources are the main reasons for poor data hygiene. First, experience is the best teacher, and few researchers have been directly responsible for data cleaning or data stewardship. Second, every study has limited resources. At the beginning of the study, the focus will invariably be on developing data collection forms and study recruitment. If data are not needed for annual reports, they may not be entered until the end of the study. The data analyst may be the first person to actually see the data. At this point, the costs required to clean the data are often greater than those required for the actual analyses.

### *Data Imputation*

At the end of the data-cleaning process, there may still be missing values that cannot be recovered. These data can be replaced using data imputation techniques. *Imputation* refers to the process of replacing a missing value with a reasonable value. Imputation methods range from mean imputation (replacing a missing data point with the average for the entire study) to *hot deck imputation* (making estimates based on a similar but complete data set), *single imputation* if the proportion of missing values is small, and *multiple imputation*. Multiple imputation remains the best technique and will be necessary if the missing data are extensive or the values are not missing at random. All methods of imputation are considered preferable to case deletion, which can result in a biased sample.

*Melinda Fritchoff Davis*

***See also*** Bias; Error; Random Error; Reliability; Systematic Error; True Score; Validity of Measurement

**Further Readings**

Allison, P. D. (2002). *Missing data*. Thousand Oaks, CA: Sage.

Davis, M. F. (2010). Avoiding data disasters and other pitfalls. In S. Sidani & D. L. Streiner (Eds.), *When research goes off the rails* (pp. 320–326). New York: Guilford.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Hoboken, NJ: Wiley.

## Data Mining

Modern researchers in various fields are confronted by an unprecedented wealth and complexity of data. However, the results available to these researchers through traditional data analysis techniques provide only limited solutions to complex situations. The approach to the huge demand for the analysis and interpretation of these complex data is managed under the name of data mining, or

*knowledge discovery. Data mining* is defined as the process of extracting useful information from large data sets through the use of any relevant data analysis techniques developed to help people make better decisions. These data mining techniques themselves are defined and categorized according to their underlying statistical theories and computing algorithms. This entry discusses these various data mining methods and their applications.

## Types of Data Mining

In general, data mining methods can be separated into three categories: *unsupervised learning, supervised learning,* and *semisupervised learning methods*. Unsupervised methods rely solely on the input variables (predictors) and do not take into account output (response) information. In unsupervised learning, the goal is to facilitate the extraction of implicit patterns and elicit the natural groupings within the data set without using any information from the output variable. On the other hand, supervised learning methods use information from both the input and output variables to generate the models that classify or predict the output values of future observations. The semisupervised method mixes the unsupervised and supervised methods to generate an appropriate classification or prediction model.

### *Unsupervised Learning Methods*

Unsupervised learning methods attempt to extract important patterns from a data set without using any information from the output variable. *Clustering analysis,* which is one of the unsupervised learning methods, systematically partitions the data set by minimizing within-group variation and maximizing between-group variation. These variations can be measured on the basis of a variety of distance metrics between observations in the data set. Clustering analysis includes hierarchical and nonhierarchical methods.

*Hierarchical clustering algorithms* provide a dendrogram that represents the hierarchical structure of clusters. At the highest level of this hierarchy is a single cluster that contains all the observations, while at the lowest level are clusters containing a single observation. Examples of hierarchical clustering algorithms are *single linkage, complete linkage, average linkage,* and *Ward's method*.

*Nonhierarchical clustering algorithms* achieve the purpose of clustering analysis without building a hierarchical structure. The *k-means clustering algorithm* is one of the most popular nonhierarchical clustering methods. A brief summary of the $k$-means clustering algorithm is as follows: Given $k$ seed (or starting) points, each observation is assigned to one of the $k$ seed points close to the observation, which creates $k$ clusters. Then seed points are replaced with the mean of the currently assigned clusters. This procedure is repeated with updated seed points until the assignments do not change. The results of the $k$-means clustering algorithm depend on the distance metrics, the number of clusters ($k$), and the location of seed points. Other nonhierarchical clustering algorithms include *k-medoids* and *self-organizing maps*.

*Principal components analysis* (PCA) is another unsupervised technique and is widely used, primarily for dimensional reduction and visualization. PCA is concerned with the covariance matrix of original variables, and the eigenvalues and eigenvectors are obtained from the covariance matrix. The product of the eigenvector corresponding to the largest eigenvalue and the original data matrix leads to the first principal component (PC), which expresses the maximum variance of the data set. The second PC is then obtained via the eigenvector corresponding to the second largest eigenvalue, and this process is repeated $N$ times to obtain $N$ PCs, where $N$ is the number of variables in the data set. The PCs are uncorrelated to each other, and generally the first few PCs are sufficient to account for most of the variations. Thus, the PCA plot of observations using these first few PC axes facilitates visualization of high-dimensional data sets.

### *Supervised Learning Methods*

Supervised learning methods use both the input and output variables to provide the model or rule that characterizes the relationships between the input and output variables. Based on the characteristics of the output variable, supervised learning methods can be categorized as either *regression* or *classification*. In regression problems, the output variable is continuous, so the main goal is to predict the outcome values of an unknown future

observation. In classification problems, the output variable is categorical, and the goal is to assign existing labels to an unknown future observation.

Linear regression models have been widely used in regression problems because of their simplicity. Linear regression is a parametric approach that provides a linear equation to examine relationships of the mean response to one or to multiple input variables. Linear regression models are simple to derive, and the final model is easy to interpret. However, the parametric assumption of an error term in linear regression analysis often restricts its applicability to complicated multivariate data. Further, linear regression methods cannot be employed when the number of variables exceeds the number of observations. *Multivariate adaptive regression spline* (MARS) is a nonparametric regression method that compensates for limitation of ordinary regression models. MARS is one of the few tractable methods for high-dimensional problems with interactions, and it estimates a completely unknown relationship between a continuous output variable and a number of input variables. MARS is a data-driven statistical linear model in which a forward stepwise algorithm is first used to select the model term and is then followed by a backward procedure to prune the model. The approximation bends at "knot" locations to model curvature, and one of the objectives of the forward stepwise algorithm is to select the appropriate knots. Smoothing at the knots is an option that may be used if derivatives are desired.

Classification methods provide models to classify unknown observations according to the existing labels of the output variable. Traditional classification methods include *linear discriminant analysis* (LDA) and *quadratic discriminant analysis* (QDA), based on Bayesian theory. Both LDA and QDA assume that the data set follows normal distribution. LDA generates a linear decision boundary by assuming that populations of different classes have the same covariance. QDA, on the other hand, does not have any restrictions on the equality of covariance between two populations and provides a quadratic equation that may be efficient for linearly nonseparable data sets.

Many supervised learning methods can handle both regression and classification problems, including decision trees, support vector machines, *k*-nearest neighbors, and artificial neural networks. *Decision tree models* have gained huge popularity in various areas because of their flexibility and interpretability. Decision tree models are flexible in that the models can efficiently handle both continuous and categorical variables in the model construction. The output of decision tree models is a hierarchical structure that consists of a series of if–then rules to predict the outcome of the response variable, thus facilitating the interpretation of the final model. From an algorithmic point of view, the decision tree model has a forward stepwise procedure that adds model terms and a backward procedure for pruning, and it conducts variable selection by including only useful variables in the model. *Support vector machine* (SVM) is another supervised learning model popularly used for both regression and classification problems. SVMs use geometric properties to obtain a separating hyperplane by solving a convex optimization problem that simultaneously minimizes the generalization error and maximizes the geometric margin between the classes. Nonlinear SVM models can be constructed from kernel functions that include linear, polynomial, and radial basis functions. Another useful supervised learning method is *k-nearest neighbors* (kNNs). A type of *lazy-learning* (*instance-based learning*) technique, kNNs do not require a trained model. Given a query point, the $k$ closest points are determined. A variety of distance measures can be applied to calculate how close each point is to the query point. Then the $k$ nearest points are examined to find which of the categories belong to the $k$ nearest points. Last, this category is assigned to the query point being examined. This procedure is repeated for all the points that require classification. Finally, *artificial neural networks* (ANNs), inspired by the way biological nervous systems learn, are widely used for prediction modeling in many applications. ANN models are typically represented by a network diagram containing several layers (e.g., input, hidden, and output layers) that consist of nodes. These nodes are interconnected with weighted connection lines whose weights are adjusted when training data are presented to the ANN during the training process. The neural network training process is an iterative adjustment of the internal weights to bring the network's output

closer to the desired values through minimizing the mean squared error.

### *Semisupervised Learning Methods*

Semisupervised learning approaches have received increasing attention in recent years. Olivier Chapelle and his coauthors described semisupervised learning as "halfway between supervised and unsupervised learning" (p. 4). Semisupervised learning methods create a classification model by using partial information from the labeled data. One-class classification is an example of a semisupervised learning method that can distinguish between the class of interest (target) and all other classes (outlier). In the construction of the classifiers, one-class classification techniques require only the information from the target class. The applications of one-class classification include *novelty detection, outlier detection,* and *imbalanced classification.*

*Support vector data description* (SVDD) is a one-class classification method that combines a traditional SVM algorithm with a density approach. SVDD produces a classifier to separate the target from the outliers. The decision boundary of SVDD is constructed from an optimization problem that minimizes the volume of the hypersphere from the boundary and maximizes the target data being captured by the boundary. The main difference between the supervised and semisupervised classification methods is that the former generates a classifier to classify an unknown observation into the predefined classes, whereas the latter gives a closed-boundary around the target data in order to separate them from all other types of data.

## Applications

Interest in data mining has increased greatly because of the availability of new analytical techniques with the potential to retrieve useful information or knowledge from vast amounts of complex data that were heretofore unmanageable. Data mining has a range of applications, including manufacturing, marketing, telecommunication, health care, biomedicine, e-commerce, and sports. In manufacturing, data mining methods have been applied to predict the number of product defects in a process and identify their causes. In marketing, market basket analysis provides a way to understand the behavior of profitable customers by analyzing their purchasing patterns. Further, unsupervised clustering analyses can be used to segment customers by market potential. In the telecommunication industries, data mining methods help sales and marketing people establish loyalty programs, develop fraud detection modules, and segment markets to reduce revenue loss. Data mining has received tremendous attention in the field of bioinformatics, which deals with large amounts of high-dimensional biological data. Data mining methods combined with *microarray technology* allow monitoring of thousands of genes simultaneously, leading to a greater understanding of molecular patterns. Clustering algorithms use microarray gene expression data to group the genes based on their level of expression, and classification algorithms use the labels of experimental conditions (e.g., disease status) to build models to classify different experimental conditions.

## Data Mining Software

A variety of data mining software is available. SAS Enterprise Miner (www.sas.com), SPSS (an IBM company, formerly called PASW® Statistics) Clementine (www.spss.com), and S-PLUS Insightful Miner (www.insightful.com) are examples of widely used commercial data mining software. In addition, commercial software developed by Salford Systems (www.salford-systems.com) provides CART, MARS, TreeNet, and Random Forests for specialized uses of tree-based models. Free data mining software packages also are available. These include RapidMiner (rapid-i.com), Weka (www.cs.waikato.ac.nz/ml/weka), and R (www.r-project.org).

*Seoung Bum Kim and Thuntee Sukchotrat*

***See also*** Exploratory Data Analyis; Exploratory Factor Analysis; Ex Post Facto Study

### Further Readings

Chapelle, O., Zien, A., & Schölkopf, B. (Eds.). (2006). *Semi-supervised learning.* Cambridge: MIT Press.

Duda, R. O., Hart, P. E., & Storl, D. G. (2001). *Pattern classification* (2nd ed.). New York: Wiley.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning*. New York: Springer.
Mitchell, T. M. (1997). *Machine learning*. New York: McGraw-Hill.
Tax, D. M. J., & Duin, R. P. W. (2004). Support vector data description. *Machine Learning, 54,* 45–66.

# Data Snooping

The term *data snooping,* sometimes also referred to as *data dredging* or *data fishing,* is used to describe the situation in which a particular data set is analyzed repeatedly without an a priori hypothesis of interest. The practice of data snooping, although common, is problematic because it can result in a significant finding (e.g., rejection of a null hypothesis) that is nothing more than a chance artifact of the repeated analyses of the data. The biases introduced by data snooping increase the more a data set is analyzed in the hope of a significant finding. Empirical research that is based on experimentation and observation has the potential to be impacted by data snooping.

## Data Snooping and Multiple Hypothesis Testing

A hypothesis test is conducted at a significance level, denoted $\alpha$, corresponding to the probability of incorrectly rejecting a true null hypothesis (the so-called Type 1 error). Data snooping essentially involves performing a large number of hypothesis tests on a particular data set with the hope that one of the tests will be significant. This data-snooping process of performing a large number of hypothesis tests results in the actual significance level being increased, or the burden of proof for finding a significant result being substantially reduced, resulting in potentially misleading results. For example, if 100 independent hypothesis tests are conducted on a data set at a significance level of 5%, it would be expected that about 5 out of the 100 tests would yield significant results simply by chance alone, even if the null hypothesis were, in fact, true. Any conclusions of statistical significance at the 5% level based on an analysis such as this are misleading because the data-snooping process has essentially ensured that something significant will be found. This means that if new data are obtained, it is unlikely that the "significant" results found via the data-snooping process would be replicated.

## Data-Snooping Examples

### *Example 1*

An investigator obtains data to investigate the impact of a treatment on the mean of a response variable of interest without a predefined view (alternative hypothesis) of the direction (positive or negative) of the possible effect of the treatment. Data snooping would occur in this situation if after analyzing the data, the investigator observes that the treatment appears to have a negative effect on the response variable and then uses a one-sided alternative hypothesis corresponding to the treatment having a negative effect. In this situation, a two-sided alternative hypothesis, corresponding to the investigator's a priori ignorance on the effect of the treatment, would be appropriate. Data snooping in this example results in the $p$ value for the hypothesis test being halved, resulting in a greater chance of assessing a significant effect of the treatment. To avoid problems of this nature, many journals require that two-sided alternatives be used for hypothesis tests.

### *Example 2*

A data set containing information on a response variable and six explanatory variables is analyzed, without any a priori hypotheses of interest, by fitting each of the 64 multiple linear regression models obtained by means of different combinations of the six explanatory variables, and then only statistically significant associations are reported. The effect of data snooping in this example would be more severe than in Example 1 because the data are being analyzed many more times (more hypothesis tests are performed), meaning that one would expect to see a number of significant associations simply due to chance.

### Correcting for Data Snooping

The ideal way to avoid data snooping is for an investigator to verify any significant results found via a data-snooping process by using an independent data set. Significant results not replicated on the independent data set would then be viewed as spurious results that were likely an artifact of the data-snooping process. If an independent data set is obtainable, then the initial data-snooping process may be viewed as an initial exploratory analysis used to inform the investigator of hypotheses of interest. In cases in which an independent data set is not possible or very expensive, the role of an independent data set can be mimicked by randomly dividing the original data into two smaller data sets: one half for an initial exploratory analysis (the *training set*) and the other half for validation (the *validation set*). Due to prohibitive cost and/or time, obtaining an independent data set or a large enough data set for dividing into training and validation sets may not be feasible. In such situations, the investigator should describe exactly how the data were analyzed, including the number of hypothesis tests that were performed in finding statistically significant results, and then report results that are adjusted for multiple hypothesis-testing effects. Methods for adjusting for multiple hypothesis testing include the *Bonferroni correction, Scheffés method, Tukey's test,* and more recently the *false discovery rate*. The relatively simple Bonferroni correction works by conducting individual hypothesis tests at level of significance $\alpha/g$ where $g$ is the number of hypothesis tests carried out. Performing the individual hypothesis tests at level of significance $\alpha/g$ provides a crude means of maintaining an overall level of significance of at least $\alpha$. Model averaging methods that combine information from every analysis of a data set are another alternative for alleviating the problems of data snooping.

*Data mining,* a term used to describe the process of exploratory analysis and extraction of useful information from data, is sometimes confused with data snooping. Data snooping is sometimes the result of the misuse of data-mining methods, such as the framing of specific alternative hypotheses in response to an observation arising out of data mining.

*Michael A. Martin and Steven Roberts*

*See also* Bonferroni Procedure; Data Mining; Hypothesis; Multiple Comparison Tests; *p* Value; Significance Level, Interpretation and Construction; Type I Error

#### Further Readings

Freedman, D., Pisani, R., & Purves, R. (2007). *Statistics* (4th ed.). New York: W. W. Norton.

Romano, J. P., & Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrika, 73,* 1237–1282.

Strube, M. J. (2006). SNOOP: A program for demonstrating the consequences of premature and repeated null hypothesis testing. *Behavior Research Methods, 38,* 24–27.

White, H. (2000). A reality check for data snooping. *Econometrika, 68,* 1097–1126.

## Debriefing

Debriefing is the process of giving participants further information about a study in which they participated at the conclusion of their participation. Debriefing continues the informational process that began at the participant recruitment or informed consent stage. If the true purpose of the study was revealed to participants at the informed consent stage, debriefing is fairly straightforward. Participants are reminded of the purposes of the study, given further information about expected results, and thanked for their participation. The debriefing session also provides an opportunity for participants to ask any questions they may have about the study. In some research situations, participants might be called on to discuss negative emotions or reveal sensitive information (e.g., studies on relationship violence or eating disorders). In such studies, the researcher may include in the debriefing information about ways in which participants might obtain help in dealing with these issues, such as a referral to a campus mental health center. A debriefing script should be included in research proposals submitted to an institutional review board.

If a study includes deception, debriefing is more complex. In such instances, a researcher has concluded that informing participants of the nature of the study at the stage of obtaining consent would interfere with the collection of valid and

generalizable data. In such instances, the researcher may give participants incomplete or misleading information about the nature of the study at the recruitment and consent stages. Other examples of deception in social science research include deceptive instructions, false feedback, or the use of confederates (members of the research team who misrepresent their identities as part of the study procedure).

In a deception study, the debriefing session is the time when a complete explanation of the study is given and the deception is revealed. Participants should be informed of the deception that took place and of the true purpose of the research. The reasons the researcher believed that deception was necessary for the research should also be explained to participants. As in a nondeception study, participants should be thanked for their participation and provided with an opportunity to ask questions of the researcher. Participants should also be reminded of their right to withdraw from the study at any time. This reminder may take a number of forms, ranging from a statement in the debriefing script indicating participants' ability to withdraw, to a second informed consent form for participants to sign after being debriefed.

## The Debriefing Process

David Holmes has argued that debriefing should include processes of *dehoaxing* (if necessary) and *desensitizing*. Dehoaxing involves informing participants about any deception that was used in the study and explaining the researcher's rationale for the use of deception. Desensitizing involves discussing and attempting to diminish any negative feelings (such as stress or anxiety) that may have arisen as a result of the research process.

Negative feelings may result from the research process for a number of reasons. The purpose of the research may have been to study these feelings, and thus researchers may have deliberately instigated them in participants. For example, researchers interested in the effects of mood on test performance might ask participants to read an upsetting passage before completing a test. Negative feelings may also arise as a consequence of engaging in the behavior that researchers were interested in studying. For example, researchers interested in conformity and compliance to authority may create situations in which participants are expected to engage in behavior with which they are uncomfortable (such as administering supposed electric shocks to a confederate). In such a situation, a researcher might address possible negative feelings by stating that the participant's behavior was not unusual or extreme (by, for example, stating that most other participants have acted the same way). Another approach is to emphasize that the behavior was due to situational factors rather than personal characteristics. Desensitizing may encourage participants to make an external (situational) rather than an internal (personal) attribution for their behavior. Participants may feel angry, foolish, or embarrassed about having been deceived by the researcher. One desensitizing technique applicable to such situations is to point out that negative feelings are a natural and expected outcome of the study situation.

Joan Seiber states that participation in research and postresearch debriefing should provide participants with new insight into the topic of research and a feeling of satisfaction in having made a contribution to society and to scientific understanding. In a deceptive study, Seiber states, participants should receive a number of additional benefits from the debriefing: dehoaxing, desensitizing, an opportunity to ask questions of the researcher, an opportunity to end participation in the study, restoration of confidence in scientific research, and information on the ways in which possible harm has been anticipated and avoided. Seiber also states that the dehoaxing process should include a convincing demonstration of the deception (for example, showing participants two identical completed tasks, one with positive feedback and one with negative feedback).

## Types of Debriefing

Several types of debriefing are associated with deception studies. In each type, the researcher describes the deceptive research processes, explains the reasons research is conducted on this topic and why deception was felt necessary to conduct the research, and thanks the participant for his or her assistance in conducting the research.

An *explicit* or *outcome debriefing* focuses on revealing the deception included in the study. Explicit debriefing would include a statement

about the deceptive processes. Explicit debriefing might also include a concrete demonstration of the deception, such as demonstrating how feedback was manipulated or introducing the participant to the confederate.

A *process debriefing* is typically more involved than an explicit debriefing and allows for more opportunities for participants to discuss their feelings about participation and reach their own conclusions regarding the study. A process debriefing might include a discussion of whether the participant found anything unusual about the research situation. The researcher might then introduce information about deceptive elements of the research study, such as false feedback or the use of confederates. Some process debriefings attempt to lead the participant to a realization of the deception on his or her own, before it is explicitly explained by the researcher.

A somewhat less common type of debriefing is an *action debriefing*, which includes an explicit debriefing along with a reenactment of the study procedure or task.

## Ethical Considerations

Ethical considerations with any research project typically include an examination of the predicted costs (e.g., potential harms) and benefits of the study, with the condition that research should not be conducted unless predicted benefits significantly outweigh potential harms. One concern expressed by ethicists is that the individuals who bear the risks of research participation (study participants) are often not the recipients of the study's benefits. Debriefing has the potential to ameliorate costs (by decreasing discomfort and negative emotional reactions) and increase benefits (by giving participants a fuller understanding of the importance of the research question being examined and thus increasing the educational value of participation).

Some experts believe that debriefing cannot be conducted in such a way as to make deception research ethical, because deceptive research practices eliminate the possibility for truly informed consent. Diana Baumrind has argued that debriefing is insufficient to remediate the potential harm caused by deception and that research involving intentional deception is unethical and should not be conducted. Other arguments against debriefing after deception include the potential for debriefing to exacerbate harm by emphasizing the deceptiveness of researchers or for participants not to believe the debriefing, inferring that it is still part of the experimental manipulation.

Other experts have argued that it is possible to conduct deception research ethically but have expressed concerns regarding possible negative outcomes of such research. One such concern regarding deception and debriefing is the *perseverance phenomenon,* in which participants continue even after debriefing to believe or be affected by false information presented in a study. The most prominent study of the perseverance phenomenon was conducted by Lee Ross, Mark Lepper, and Michael Hubbard, who were interested in adolescents' responses to randomly assigned feedback regarding their performance on a decision-making task. At the end of the study session, participants participated in a debriefing session in which they learned that the feedback they had received was unrelated to their actual performance. Ross and colleagues found that participants' self-views were affected by the feedback even after the debriefing. When, as part of the debriefing process, participants were explicitly told about the perseverance phenomenon, their self-views did not continue to be affected after the debriefing.

## Debriefing in Particular Research Contexts

Most of the preceding discussion of debriefing has assumed a study of adult participants in a laboratory setting. Debriefing may also be used in other research contexts, such as Internet research, or with special research populations, such as children or members of stigmatized groups. In Internet research, debriefing is typically presented in the form of a debriefing statement as the final page of the study or as an e-mail sent to participants.

In research with children, informed consent prior to participation is obtained from children's parents or guardians; child participants give their assent as well. In studies involving deception of child participants, parents are typically informed of the true nature of the research at the informed consent stage but are asked not to reveal the nature of the research project to their children prior to participation in the study. After study participation, children participate in a debriefing

session with the researcher (and sometimes with a parent or guardian as well). In this session, the researcher explains the nature of and reasons for the deception in age-appropriate language.

Marion Underwood has advocated for the use of a process debriefing with children. Underwood has also argued that it is important for the deception and debriefing to take place within a larger context of positive interactions. For example, children might engage in an enjoyable play session with a child confederate after being debriefed about the confederate's role in an earlier interaction.

### Statements by Professional Organizations

The American Psychological Association's Ethical Principles of Psychologists and Code of Conduct states that debriefing should be an opportunity for participants to receive appropriate information about a study's aims and conclusions and should include correction of any participant misperceptions of which the researchers are aware. The APA's ethics code also states that if information must be withheld for scientific or humanitarian reasons, researchers should take adequate measures to reduce the risk of harm. If researchers become aware of harm to a participant, they should take necessary steps to minimize the harm. The ethics code is available on the APA's Web site.

The Society for Research in Child Development's Ethical Standards for Research with Children state that the researcher should clarify all misconceptions that may have arisen over the course of the study immediately after the data are collected. This ethics code is available on the Society's Web site.

*Meagan M. Patterson*

***See also*** Ethics in the Research Process; Informed Consent

#### Further Readings

Baumrind, D. (1985). Research using intentional deception: Ethical issues revisited. *American Psychologist, 40,* 165–174.

Elms, A. (1982). Keeping deception honest: Justifying conditions for social scientific research strategems. In T. L. Beauchamp, R. R. Faden, R. J. Wallace Jr., 245). Baltimore: Johns Hopkins University Press.

Fisher, C. B. (2005). Deception research involving children: Ethical practices and paradoxes. *Ethics 287.*

Holmes, D. S. (1976). Debriefing after psychological experiments: I. Effectiveness of postdeception dehoaxing. *American Psychologist, 31,* 868–875.

Hurley, J. C., & Underwood, M. K. Children's understanding of their research rights before and after debriefing: Informed assent, confidentiality, and stopping participation. *Child Development, 73,* 132–143.

Mills, J. (1976). A procedure for explaining experiments involving deception. *Personality 13.*

Sieber, J. E. (1983). Deception in social research III: The nature and limits of debriefing. *IRB: A Review of Human Subjects Research, 5,* 1–4.

#### Websites

American Psychological Association: http://www.apa.org

Society for Research in Child Development's Ethical Standards for Research with Children: http://www.srcd.org

## Decision Rule

In the context of statistical hypothesis testing, decision rule refers to the rule that specifies how to choose between two (or more) competing hypotheses about the observed data. A decision rule specifies the statistical parameter of interest, the test statistic to calculate, and how to use the test statistic to choose among the various hypotheses about the data. More broadly, in the context of statistical decision theory, a decision rule can be thought of as a procedure for making rational choices given uncertain information.

The choice of a decision rule depends, among other things, on the nature of the data, what one needs to decide about the data, and at what level of significance. For instance, decision rules used for normally distributed (or Gaussian) data are generally not appropriate for non-Gaussian data. Similarly, decision rules used for determining the 95% confidence interval of the sample mean will be different from the rules appropriate for binary decisions, such as determining whether the sample mean is greater than a prespecified mean value at

a given significance level. As a practical matter, even for a given decision about a given data set, there is no unique, universally acceptable decision rule but rather many possible principled rules.

There are two main statistical approaches to picking the most appropriate decision rule for a given decision. The classical, or *frequentist,* approach is the one encountered in most textbooks on statistics and the one used by most researchers in their data analyses. This approach is generally quite adequate for most types of data analysis. The *Bayesian* approach is still widely considered esoteric, but one that an advanced researcher should become familiar with, as this approach is becoming increasingly common in advanced data analysis and complex decision making.

## Decision Rules in Classical Hypothesis Testing

Suppose one needs to decide whether a new brand of bovine growth hormone increases the body weight of cattle beyond the known average value of $\mu$ kilograms. The observed data consist of body weight measurements from a sample of cattle treated with the hormone. The default explanation for the data, or the *null hypothesis,* is that there is no effect: the mean weight of the treated sample is no greater than the nominal mean $\mu$. The *alternative hypothesis* is that the mean weight of the treated sample is greater than $\mu$.

The decision rule specifies how to decide which of the two hypotheses to accept, given the data. In the present case, one may calculate the $t$ statistic, determine the *critical value* of $t$ at the desired level of significance (such as .05), and accept the alternative hypothesis if the $t$ value based on the data exceeds the critical value and reject it otherwise. If the sample is sufficiently large and Gaussian, one might use a similar decision rule with a different test statistic, the $z$ score. Alternatively, one may choose between the hypotheses based on the $p$ value rather than the critical value.

Such case-specific variations notwithstanding, what all frequentist decision rules have in common is that they arrive at a decision ultimately by comparing some statistic of the observed data against a theoretical standard, such as the sampling distribution of the statistic, and determine how likely the observed data are under the various competing hypotheses.

Conceptual quibbles about this view of probability aside, this approach is entirely adequate for a vast majority of practical purposes in research. But for more complex decisions in which a variety of factors and their attendant uncertainties have to be considered, frequentist decision rules are often too limiting.

## Bayesian Decision Rules

Suppose, in the aforementioned example, that the effectiveness of the hormone for various breeds of cattle in the sample, and the relative frequencies of the breeds, is known. How should one use this *prior distribution* of hormone effectiveness to choose between the two hypotheses? Frequentist decision rules are not well suited to handle such decisions; Bayesian decision rules are.

Essentially, Bayesian decision rules use Bayes's law of conditional probability to compute a *posterior distribution* based on the observed data and the appropriate prior distribution. In the case of the above example, this amounts to revising one's belief about the body weight of the treated cattle based on the observed data and the prior distribution. The null hypothesis is rejected if the *posterior probability* is less than the user-defined significance level.

One of the more obvious advantages of Bayesian decision making, in addition to the many subtler ones, is that Bayesian decision rules can be readily elaborated to allow any number of additional considerations underlying a complex decision. For instance, if the larger decision at hand in the above example is whether to market the hormone, one must consider additional factors, such as the projected profits, possible lawsuits, and costs of manufacturing and distribution. Complex decisions of this sort are becoming increasingly common in behavioral, economic, and social research. Bayesian decision rules offer a statistically optimal method for making such decisions.

It should be noted that when only the sample data are considered and all other factors, including prior distributions, are left out, Bayesian decision rules can lead to decisions equivalent to and even identical to the corresponding frequentist rules. This superficial similarity between the two approaches notwithstanding, Bayesian decision rules are not simply a more elaborate version of frequentist rules.

The differences between the two approaches are profound and reflect longstanding debates about the nature of probability. For the researcher, on the other hand, the choice between the two approaches should be less a matter of adherence to any given orthodoxy and more about the nature of the decision at hand.

*Jay Hegdé*

*See also* Criterion Problem; Critical Difference; Error Rates; Expected Value; Inference: Inductive and Deductive; Mean Comparisons; Parametric Statistics

**Further Readings**

Bolstad, W. M. (2007). *Introduction to Bayesian statistics*. Hoboken, NJ: Wiley.
Press, S. J. (2005). *Applied multivariate analysis: Using Bayesian and frequentist methods of inference*. New York: Dover.
Resnik, M. D. (1987). *Choices: An introduction to decision theory*. Minneapolis: University of Minnesota Press.

# Declaration of Helsinki

The Declaration of Helsinki is a formal statement of ethical principles published by the World Medical Association (WMA) to guide the protection of human participants in medical research. The Declaration is not a legally binding document but has served as a foundation for national and regional laws governing medical research across the world. Although not without its controversies, the Declaration has served as the standard in medical research ethics since its establishment in 1964.

## History and Current Status

Before World War II, no formal international statement of ethical principles to guide research with human participants existed, leaving researchers to rely on organizational, regional, or national policies or their own personal ethical guidelines. After atrocities were found to have been committed by Nazi medical researchers using involuntary, unprotected participants drawn from concentration camps, the 1947 Nuremberg Code was established. This was followed in 1948 by the WMA's Declaration of Geneva, a statement of ethical duties for physicians. Both documents influenced the development of the Declaration of Helsinki, adopted in 1964 by the WMA. The initial Declaration, 11 paragraphs in length, focused on clinical research trials. Notably, it relaxed conditions for consent for participation, changing the Nuremberg requirement that consent is "absolutely essential" to instead urge consent "if at all possible" but to allow for proxy consent, such as from a legal guardian, in some instances.

The Declaration has been revised six times. The first revision, conducted in 1975, expanded the Declaration considerably, nearly doubling its length, increasing its depth, updating its terminology, and adding concepts such as oversight by an independent committee. The second (1983) and third (1989) revisions were comparatively minor, primarily involving clarifications and updates in terminology. The fourth (1996) revision also was minor in scope but notably added a phrase that effectively precluded the use of inert placebos when a particular standard of care exists.

The fifth (2000) revision was extensive and controversial. In the years leading up to the revision, concerns were raised about the apparent use of relaxed ethical standards for clinical trials in developing countries, including the use of placebos in HIV trials conducted in sub-Saharan Africa. Debate ensued about revisions to the Declaration, with some arguing for stronger language and commentary addressing clinical trials and others proposing to limit the document to basic guiding principles. Although consensus was not reached, the WMA approved a revision that restructured the document and expanded its scope. Among the more controversial aspects of the revision was the implication that standards of medical care in developed countries should apply to any research with humans, including that conducted in developing countries. The opposing view held that when risk of harm is low and there are no local standards of care (as is often the case in developing countries), placebo-controlled trials are ethically acceptable, especially given their potential benefits for future patients. Debate has continued on these issues, and cross-national divisions have emerged. The

U.S. Food and Drug Administration rejected the fifth revision because of its restrictions on the use of placebo conditions and has eliminated all references to the Declaration, replacing it with the Good Clinical Practice guidelines, an alternative internationally sanctioned ethics guide. The National Institutes of Health training in research with human participants no longer refers to the Declaration, and the European Commission refers only to the fourth revision.

The sixth revision of the Declaration, approved by the WMA in 2008, introduced relatively minor clarifications. The revision reinforces the Declaration's long-held emphasis on prioritizing the rights of individual research participants above all other interests. Public debate following the revision was not nearly as contentious as had been the case with previous revisions.

## Synopsis of the Sixth Revision

The Declaration of Helsinki's sixth revision comprises several sections: the Introduction, Principles for All Medical Research, and Additional Principles for Medical Research Combined With Medical Care. It is 35 paragraphs long.

### *Introduction*

The introduction states that the Declaration is intended for physicians and others who conduct medical research on humans (including human materials or identifiable information). It asserts that the Declaration should be considered as a whole and that its paragraphs should not be considered in isolation but with reference to all pertinent paragraphs. It then outlines general ethical principles that guide research on human participants. These include a reminder of the words from the WMA's Declaration of Geneva that the physician is bound to: "The health of my patient will be my first consideration." This idea is expanded with a statement asserting that when research is being conducted, the welfare of the participants takes precedence over the more general welfare of science, research, and the general population.

The introduction also describes the goals of medical research as improving the prevention, diagnosis, and treatment of disease and increasing the understanding of the etiology of disease. It adds that research also must encourage the protection of the health and rights of people. The introduction then specifically mentions vulnerable populations and calls for extra consideration when these populations are participating in research. The final statement in the Declaration's Introduction asserts that medical researchers are bound by the legal and ethical guidelines of their own nations but that adherence to these laws does not liberate researchers from the edicts of the Declaration of Helsinki.

### *Principles for All Medical Research*

The Principles for All Medical Research include considerations that must be made by researchers who work with human participants. The first assertion in the principles states that a physician's duty is to "protect the life, health, dignity, integrity, right to self-determination, privacy, and confidentiality" of research participants. Consideration of the environment and of the welfare of research animals is also mentioned. Also, the basic principles declare that any research conducted on human participants must be in accordance with generally held scientific principles and be based on as thorough a knowledge of the participant as is possible.

Paragraph 14 of the Declaration states that any study using human participants must be thoroughly outlined in a detailed protocol, and it provides specific guidelines about what should be included in the protocol. The protocol should include numerous types of information, including funding sources, potential conflicts of interest, plans for providing study participants access to interventions that the study identifies as beneficial, and more.

Paragraph 15 states that the above-mentioned protocol must be reviewed by an independent research ethics committee before the study begins. This committee has the right and responsibility to request changes, provide comments and guidance, and monitor ongoing trials. The committee members also have the right and responsibility to consider all information provided in the protocol and to request additional information as deemed appropriate. This principle of the Declaration is what has led to the development of institutional review boards in the United States.

The principles also state that research must be conducted by qualified professionals and that the responsibility for protecting research subjects always falls on the professionals conducting the study and not on the study participants, even though they have consented to participate.

The principles also require an assessment of predictable risks and benefits for both research participants and the scientific community. Risk management must be carefully considered, and the objective of the study must be of enough importance that the potential risks are outweighed. Another statement in the basic principles, paragraph 17, states that research with disadvantaged or vulnerable populations is justified only if it relates to the needs and priorities of the vulnerable community and can be reasonably expected to benefit the population in which the research is conducted. This statement was included, in part, as a response to testing of new prescription drugs in Africa, where the availability of cutting-edge prescription drugs is highly unlikely.

The remainder of the principles section discusses issues of privacy, confidentiality, and informed consent. These discussions stipulate that research should be conducted only with participants who are capable of providing informed consent, unless it is absolutely necessary to do research with participants who cannot give consent. If this is the case, the specific reasons for this necessity must be outlined in the protocol, informed consent must be provided by a legal guardian, and the research participant's assent must be obtained if possible. Participants must be informed of their right to refuse to participate in the study, and special care must be taken when potential participants are under the care of a physician involved in the study in order to avoid dynamics of dependence on the physician or duress to affect decision-making processes. Paragraphs 27 through 29 outline guidelines for research with participants who are deemed incompetent to give consent and state that these subjects can be included in research only if the subject can be expected to benefit or if the following conditions apply: A population that the participant represents is likely to benefit, the research cannot be performed on competent persons, and potential risk and burden are minimal. The final paragraph of the principles addresses the ethics of accurate publication of research results. Researchers are responsible for accurate and complete reporting of results and for making their results publicly available, even if the results are negative or inconclusive. The publication should also include funding sources, institutional affiliation, and any conflicts of interest. A final assertion states that research reports that do not meet these standards should not be accepted for publication.

### *Additional Principles for Medical Research Combined With Medical Care*

This section of the Declaration, which was new to the fifth revision in 2000, has created the most controversy. It begins with a statement that extra care must be taken to safeguard the health and rights of patients who are both receiving medical care and participating in research. Paragraph 32 then states that when a new treatment method is being tested, it should be compared with the generally accepted best standard of care, with two exceptions. First, placebo treatment can be used in studies where no scientifically proven intervention exists. This statement was adopted as a response to drug testing that was being conducted in which the control group was given placebos when a scientifically proven drug was available.

The second exception states that placebos or no treatment can be used when "compelling and scientifically sound methodological reasons" exist for using a placebo to determine the efficacy and/or safety of a treatment, and if the recipients of the placebo or no treatment will not suffer irreversible harm. The Declaration then states that "Extreme care must be taken to avoid abuse of this option." This exception was most likely added as a response to the intense criticism of the fifth revision.

The adoption of the principle described in paragraph 32 aimed to prevent research participants' illnesses from progressing or being transmitted to others because of a lack of drug treatment when a scientifically proven treatment existed. Critics of this assertion stated that placebo treatment was consistent with the standard of care in the regions where the drug testing was taking place and that administration of placebos to control groups is often necessary to determine the efficacy of

a treatment. Supporters of the Declaration maintain that the duty of the medical professional is to provide the best care possible to patients and that knowingly administering placebos in place of proven treatments is ethically dubious.

Paragraph 33 of the Declaration establishes that study participants "are entitled to be informed about the outcome of the study and to share any benefits that result from it" and gives the example of participants' being provided access to interventions that have been identified as beneficial, or to other appropriate care. This assertion is somewhat less strong than the fifth edition's language, which stated that participants should be "assured" of the best known care identified during the course of the study when a study is concluded.

The final two paragraphs of the Declaration, which are part of the "Additional Principles," provide that a patient's refusal to participate in a study should never affect the therapeutic relationship and that new, unproven treatments can be used when there is reason to believe they will be beneficial and when no proven treatment exists.

### Future

The Declaration of Helsinki remains the world's best-known statement of ethical principles to guide medical research with human participants. Its influence is far-reaching in that it has been codified into the laws that govern medical research in countries across the world and has served as a basis for the development of other international guidelines governing medical research with human participants. As the Declaration has expanded and become more prescriptive, it has become more controversial, and concerns have been raised regarding the future of the Declaration and its authority. Future revisions to the Declaration may reconsider the utility of prescriptive guidelines rather than limiting its focus to basic principles. Another challenge will be to harmonize the Declaration with other ethical research guidelines, because there often is apparent conflict between aspects of current codes and directives documents.

*Bryan J. Dik and Timothy J. Doenges*

***See also*** Ethics in the Research Process

**Further Readings**

Carlson, R. V., van Ginneken, N. H., Pettigrew, L. M., Davies, A., Boyd, K. M., & Webb, D. J. (2007). The three official language versions of the Declaration of Helsinki: What's lost in translation? *Journal of Medical Ethics, 33,* 545–548.

Human, D., & Fluss, S. S. (2001). *The World Medical Association's Declaration of Helsinki: Historical and contemporary perspectives*. Retrieved February 1, 2010, from http://www.wma.net/en/20activities/10ethics/10helsinki/draft_historical_contemporary_perspectives.pdf

Schmidt, U., & Frewer, A. (Eds.). (2007). *History and theory of human experimentation: The Declaration of Helsinki and modern medical ethics*. Stuttgart, Germany: Franz Steiner.

Williams, J. R. (2008). The Declaration of Helsinki and public health. *Bulletin of the World Health Organization, 86,* 650–651.

Williams, J. R. (2008). Revising the Declaration of Helsinki. *World Medical Journal, 54,* 120–125.

Wolinsky, H. (2006). The battle of Helsinki. *EMBO Reports, 7,* 670–672.

World Medical Association. (2008). *The Declaration of Helsinki*. Retrieved February 1, 2010, from http://www.wma.net/en/30publications/10policies/b3/index.html

## Degrees of Freedom

In statistics, the degrees of freedom is a measure of the level of precision required to estimate a parameter (i.e., a quantity representing some aspect of the population). It expresses the number of independent factors on which the parameter estimation is based and is often a function of sample size. In general, the number of degrees of freedom increases with increasing sample size and with decreasing number of estimated parameters. The quantity is commonly abbreviated *df* or denoted by the lowercase Greek letter *nu, ν*.

For a set of observations, the degrees of freedom is the minimum number of independent values required to resolve the entire data set. It is equal to the number of independent observations being used to determine the estimate (*n*) minus the number of parameters being estimated in the approximation of the parameter itself, as determined by the statistical procedure under

consideration. In other words, a mathematical restraint is used to compensate for estimating one parameter from other estimated parameters. For a single sample, one parameter is estimated. Often the population mean ($\mu$), a frequently unknown value, is based on the sample mean ($\bar{x}$), thereby resulting in $n - 1$ degrees of freedom for estimating population variability. For two samples, two parameters are estimated from two independent samples ($n_1$ and $n_2$), thus producing $n_1 + n_2 - 2$ degrees of freedom. In simple linear regression, the relationship between two variables, $x$ and $y$, is described by the equation $y = bx + a$, where $b$ is the slope of the line and $a$ is the $y$-intercept (i.e., where the line crosses the $y$-axis). In estimating $a$ and $b$ to determine the relationship between the independent variable $x$ and dependent variable $y$, 2 degrees of freedom are then lost. For multiple sample groups ($n_1 + \cdots + n_k$), the number of parameters estimated increases by $k$, and subsequently, the degrees of freedom is equal to $n_1 + \cdots + n_k - k$. The denominator in the analysis of variance (ANOVA) $F$ test statistic, for example, accounts for estimating multiple population means for each group under comparison.

The concept of degrees of freedom is fundamental to understanding the estimation of population parameters (e.g., mean) based on information obtained from a sample. The amount of information used to make a population estimate can vary considerably as a function of sample size. For instance, the standard deviation (a measure of variability) of a population estimated on a sample size of 100 is based on 10 times more information than is a sample size of 10. The use of large amounts of independent information (i.e., a large sample size) to make an estimate of the population usually means that the likelihood that the sample estimates are truly representative of the entire population is greater. This is the meaning behind the number of degrees of freedom. The larger the degrees of freedom, the greater the confidence the researcher can have that the statistics gained from the sample accurately describe the population.

To demonstrate this concept, consider a sample data set of the following observations ($n = 5$): 1, 2, 3, 4, and 5. The sample mean (the sum of the observations divided by the number of observations) equals 3, and the deviations about the mean are $-2$, $-1$, 0, $+1$, and $+2$, respectively. Since the sum of the deviations about the mean is equal to zero, at least four deviations are needed to determine the fifth; hence, one deviation is fixed and cannot vary. The number of values that are free to vary is the degrees of freedom. In this example, the number of degrees of freedom is equal to 4; this is based on five data observations ($n$) minus one estimated parameter (i.e., using the sample mean to estimate the population mean). Generally stated, the degrees of freedom for a single sample are equal to $n - 1$ given that if $n - 1$ observations and the sample mean are known, the remaining $n$th observation can be determined.

Degrees of freedom are also often used to describe assorted data distributions in comparison with a normal distribution. Used as the basis for statistical inference and sampling theory, the normal distribution describes a data set characterized by a bell-shaped probability density function that is symmetric about the mean. The chi-square distribution, applied usually to test differences among proportions, is positively skewed with a mean defined by a single parameter, the degrees of freedom. The larger the degrees of freedom, the more the chi-square distribution approximates a normal distribution. Also based on the degrees of freedom parameter, the Student's $t$ distribution is similar to the normal distribution, but with more probability allocated to the tails of the curve and less to the peak. The largest difference between the $t$ distribution and the normal occurs for degrees of freedom less than about 30. For tests that compare the variance of two or more populations (e.g., ANOVA), the positively skewed $F$ distribution is defined by the number of degrees of freedom for the various samples under comparison.

Additionally, George Ferguson and Yushio Takane have offered a geometric interpretation of degrees of freedom whereby restrictions placed on the statistical calculations are related to a point–space configuration. Each point within a space of $d$ dimensions has a freedom of movement or variability within those $d$ dimensions that is equal to $d$; hence, $d$ is the number of degrees of freedom. For instance, a data point on a single dimensional line has one degree of movement (and one degree of freedom) whereas a data point in three-dimensional space has three.

*Jill S. M. Coleman*

*See also* Analysis of Variance (ANOVA); Chi-Square Test; Distribution; ***F*** Test; Normal Distribution; Parameters; Population; Sample Size; Student's *t* Test; Variance

**Further Readings**

Ferguson, G. A., & Takane, Y. (1989). *Statistical analysis in psychology and education* (6th ed.). New York: McGraw-Hill.

Good, I. J. (1973). What are degrees of freedom? *American Statistician, 27,* 227–228.

Lomax, R. G. (2001). *An introduction to statistical concepts for education and the behavioral sciences.* Mahwah, NJ: Lawrence Erlbaum.

# Delphi Technique

The Delphi technique is a group communication process as well as a method of achieving a consensus of opinion associated with a specific topic. Predicated on the rationale that more heads are better than one and that inputs generated by experts based on their logical reasoning are superior to simply guessing, the technique engages a group of identified experts in detailed examinations and discussions on a particular issue for the purpose of policy investigation, goal setting, and forecasting future situations and outcomes. Common surveys try to identify what is. The Delphi technique attempts to assess what could or should be.

The Delphi technique was named after the oracle at Delphi, who, according to Greek myth, delivered prophecies. As the name implies, the Delphi technique was originally developed to forecast future events and possible outcomes based on inputs and circumstances. The technique was principally developed by Norman Dalkey and Olaf Helmer at the RAND Corporation in the early 1950s. The earliest use of the Delphi process was primarily military. Delphi started to gain popularity as a futuring tool in the mid-1960s and came to be widely applied and examined by researchers and practitioners in fields such as curriculum development, resource utilization, and policy determination. In the mid-1970s, however, the popularity of the Delphi technique began to decline. Currently, using the Delphi technique as an integral part or as the exclusive tool of investigation in a research or an evaluation project is not uncommon.

This entry examines the Delphi process, including subject selection and analysis of data. It also discusses the advantages and disadvantages of the Delphi technique, along with the use of electronic technologies in facilitating implementation.

## The Delphi Process

The Delphi technique is characterized by multiple iterations, or "rounds," of inquiry. The iterations mean a series of feedback processes. Due to the iterative characteristic of the Delphi technique, instrument development, data collection, and questionnaire administration are interconnected between rounds. As such, following the more or less linear steps of the Delphi process is important to success with this technique.

### *Round 1*

In Delphi, one of two approaches can be taken in the initial round. Traditionally, the Delphi process begins with an open-ended questionnaire. The open-ended questionnaire serves as the cornerstone for soliciting information from invited participants. After receiving responses from participants, investigators convert the collected qualitative data into a structured instrument, which becomes the second-round questionnaire. A newer approach is based on an extensive review of the literature. To initiate the Delphi process, investigators directly administer a structured questionnaire based on the literature and use it as a platform for questionnaire development in subsequent iterations.

### *Round 2*

Next, Delphi participants receive a second questionnaire and are asked to review the data developed from the responses of all invited participants in the first round and subsequently summarized by investigators. Investigators also provide participants with their earlier responses to compare with the new data that has been summarized and edited. Participants are then asked to rate or rank order the new statements and are encouraged to express any skepticism, questions, and justifications regarding the statements. This allows a full

and fair disclosure of what each participant thinks or believes is important concerning the issue being investigated, as well as providing participants an opportunity to share their expertise, which is a principal reason for their selection to participate in the study.

### *Round 3*

In Round 3, Delphi participants receive a third questionnaire that consists of the statements and ratings summarized by the investigators after the preceding round. Participants are again asked to revise their judgments and to express their rationale for their priorities. This round provides participants an opportunity to make further clarifications and review previous judgments and inputs from the prior round. Researchers have indicated that three rounds are often sufficient to gather needed information and that further iterations would merely generate slight differences.

### *Round 4*

However, when necessary, in the fourth and often final round, participants are again asked to review the summary statements from the preceding round and to provide inputs and justifications. It is imperative to note that the number of Delphi iterations relies largely on the degree of consensus sought by the investigators and thereby can vary from three to five. In other words, a general consensus about a noncritical topic may only require three iterations, whereas a serious issue of critical importance with a need for a high level of agreement among the participants may require additional iterations. Regardless of the number of iterations, it must be remembered that the purpose of the Delphi is to sort through the ideas, impressions, opinions, and expertise of the participants to arrive at the core or salient information that best describes, informs, or predicts the topic of concern.

## Subject Selection

The proper use of the Delphi technique and the subsequent dependability of the generated data rely in large part on eliciting expert opinions. Therefore, the selection of appropriate participants is considered the most important step in Delphi. The quality of results directly links to the quality of the participants involved.

Delphi participants should be highly trained and possess expertise associated with the target issues. Investigators must rigorously consider and examine the qualifications of Delphi subjects. In general, possible Delphi subjects are likely to be positional leaders, authors discovered from a review of professional publications concerning the topic, and people who have firsthand relationships with the target issue. The latter group often consists of individuals whose opinions are sought because their direct experience makes them a reliable source of information.

In Delphi, the number of participants is generally between 15 and 20. However, what constitutes an ideal number of participants in a Delphi study has never achieved a consensus in the literature. Andre Delbecq, Andrew Van de Ven, and David Gustafson suggest that 10 to 15 participants should be adequate if their backgrounds are similar. In contrast, if a wide variety of people or groups or a wide divergence of opinions on the topic are deemed necessary, more participants need to be involved. The number of participants in Delphi is variable, but if the number of participants is too small, they may be unable to reliably provide a representative pooling of judgments concerning the target issue. Conversely, if the number of participants is too large, the shortcomings inherent in the Delphi technique (difficulty dedicating large blocks of time, low response rates) may take effect.

## Analysis of Data

In Delphi, decision rules must be established to assemble, analyze, and summarize the judgments and insights offered by the participants. Consensus on a topic can be determined if the returned responses on that specific topic reach a prescribed or a priori range. In situations in which rating or rank ordering is used to codify and classify data, the definition of consensus has been at the discretion of the investigator(s). One example of consensus from the literature is having 80% of subjects' votes fall within two categories on a 7-point scale.

The Delphi technique can employ and collect both qualitative and quantitative information.

Investigators must analyze qualitative data if, as with many conventional Delphi studies, open-ended questions are used to solicit participants' opinions in the first round. It is recommended that a team of researchers and/or experts with knowledge of both the target issues and instrument development analyze the written comments. Statistical analysis is performed in the further iterations to identify statements that achieve the desired level of consensus. Measures of central tendency (means, mode, and median) and level of dispersion (standard deviation and interquartile range) are the major statistics used to report findings in the Delphi technique. The specific statistics used depend on the definition of consensus set by the investigators.

## Advantages of Using the Delphi Technique

Several components of the Delphi technique make it suitable for evaluation and research problems. First, the technique allows investigators to gather subjective judgments from experts on problems or issues for which no previously researched or documented information is available. Second, the multiple iterations allow participants time to reflect and an opportunity to modify their responses in subsequent iterations. Third, Delphi encourages innovative thinking, particularly when a study attempts to forecast future possibilities. Last, participant anonymity minimizes the disadvantages often associated with group processes (e.g., bandwagon effect) and frees subjects from pressure to conform. As a group communication process, the technique can serve as a means of gaining insightful inputs from experts without the requirement of face-to-face interactions. Additionally, confidentiality is enhanced by the geographic dispersion of the participants, as well as the use of electronic devices such as e-mail to solicit and exchange information.

## Limitations of the Delphi Technique

Several limitations are associated with Delphi. First, a Delphi study can be time-consuming. Investigators need to ensure that participants respond in a timely fashion because each round rests on the results of the preceding round.

Second, low response rates can jeopardize robust feedback. Delphi investigators need both a high response rate in the first iteration and a desirable response rate in the following rounds. Investigators need to play an active role in helping to motivate participants, thus ensuring as high a response rate as possible.

Third, the process of editing and summarizing participants' feedback allows investigators to impose their own views, which may impact participants' responses in later rounds. Therefore, Delphi investigators must exercise caution and implement appropriate safeguards to prevent the introduction of bias.

Fourth, an assumption regarding Delphi participants is that their knowledge, expertise, and experience are equivalent. This assumption can hardly be justified. It is likely that the knowledge bases of Delphi participants are unevenly distributed. Although some panelists may have much more in-depth knowledge of a specific, narrowly defined topic, other panelists may be more knowledgeable about a wide range of topics. A consequence of this disparity may be that participants who do not possess in-depth information may be unable to interpret or evaluate the most important statements identified by Delphi participants who have in-depth knowledge. The outcome of such a Delphi study could be a series of *general* statements rather than an in-depth exposition of the topic.

## Computer-Assisted Delphi Process

The prevalence and application of electronic technologies can facilitate the implementation of the Delphi process. The advantages of computer-assisted Delphi include participant anonymity, reduced time required for questionnaire and feedback delivery, readability of participant responses, and the easy accessibility provided by Internet connections.

If an e-mail version of the questionnaires is to be used, investigators must ensure e-mail addresses are correct, contact invited participants beforehand, ask their permission to send materials via e-mail, and inform the recipients of the nature of the research so that they will not delete future e-mail contacts. With regard to the purchase of a survey service, the degree of flexibility in questionnaire templates and software and service costs

may be the primary considerations. Also, Delphi participants need timely instructions for accessing the designated link and any other pertinent information.

*Chia-Chien Hsu and Brian A. Sandford*

*See also* Qualitative Research; Quantitative Research; Survey

### Further Readings

Adler, M., & Ziglio, E. (1996). *Gazing into the oracle: The Delphi method and its application to social policy and public health*. London: Jessica Kingsley.

Altschuld, J. W., & Thomas, P. M. (1991). Considerations in the application of a modified scree test for Delphi survey data. *Evaluation Review, 15*(2), 179–188.

Dalkey, N. C., Rourke, D. L., Lewis, R., & Snyder, D. (1972). *Studies in the quality of life: Delphi and decision-making*. Lexington, MA: Lexington Books.

Delbecq, A. L., Van de Ven, A. H., & Gustafson, D. H. (1975). *Group technique for program planning: A guide to nominal group and Delphi processes*. Glenview, IL: Scott, Foresman.

Hasson, F., Keeney, S., & McKenna, H. (2000). Research guidelines for the Delphi survey technique. *Journal of Advanced Nursing, 32*(4), 1008–1015.

Hill, K. Q., & Fowles, J. (1975). The methodological worth of the Delphi forecasting technique. *Technological Forecasting & Social Change, 7*, 179–19 2.

Hsu, C. C., & Sandford, B. A. (2007, August). The Delphi technique: Making sense of consensus. *Practical Assessment, Research, & Evaluation, 12*(10). Retrieved September 15, 2007, from http://pareonline.net/getvn.asp?v = 12&n = 10

Linstone, H. A., & Turoff, M. (1975). *The Delphi method: Techniques and applications*. Reading, MA: Addison-Wesley.

Yousuf, M. I. (2007, May). Using experts' opinions through Delphi technique. *Practical Assessment, Research, & Evaluation, 12*(4). Retrieved June 28, 2007, from http://pareonline.net/getvn.asp?v = 12&n = 4

# Demographics

The term *demographics* refers to particular characteristics of a population. The word is derived from the Greek words for *people* (*demos*) and *picture* (*graphy*). Examples of demographic characteristics include age, race, gender, ethnicity, religion, income, education, home ownership, sexual orientation, marital status, family size, health and disability status, and psychiatric diagnosis.

## Demographics as Variables in Research

Demographic information provides data regarding research participants and is necessary for the determination of whether the individuals in a particular study are a representative sample of the target population for generalization purposes. Usually demographics or research participant characteristics are reported in the methods section of the research report and serve as independent variables in the research design. Demographic variables are independent variables by definition because they cannot be manipulated. In research, demographic variables may be either categorical (e.g., gender, race, marital status, psychiatric diagnosis) or continuous (e.g., age, years of education, income, family size). Demographic information describes the study sample, and demographic variables also can be explored for their moderating effect on dependent variables.

## The Nature of Demographic Variables

Some demographic variables are necessarily categorical, such as gender, whereas other demographic variables (e.g., education, income) can be collected to yield categorical or continuous variables. For example, to have education as a continuous variable, one would ask participants to report number of years of education. But to have education as a categorical variable, one would ask participants to select a category of education (e.g., less than high school, high school, some college, college degree, graduate degree). Note that a researcher could post hoc create a categorical variable for education if the data were initially gathered to yield a continuous variable.

## Defining Demographic Variables

Researchers should clearly and concisely define the demographic variables employed in their study. When possible, variables should be defined consistent with commonly used definitions or taxonomies (e.g., U.S. Census Bureau categories of

ethnicity). It is generally agreed and advisable that demographic information should be collected on the basis of participant report and not as an observation of the researcher. In the case of race, for example, it is not uncommon for someone whom a researcher may classify as Black to self-identify as White or biracial.

### Selection of Demographic Information to Be Collected

Researchers should collect only the demographic information that is necessary for the specific purposes of the research. To do so, in the planning stage researchers will need to identify demographic information that is vital in the description of participants as well as in data analysis, and also information that will enhance interpretation of the results. For example, in a study of maternal employment and children's achievement, Wendy Goldberg and colleagues found that the demographic variables of children's age and family structure were significant moderators of the results. Thus, the inclusion of particular demographic information can be critical for an accurate understanding of the data.

### Confidentiality

Respondents should be informed that demographic information will be held in strictest confidence and reported only as aggregated characteristics, not as individual data, and that the information will be used for no other purpose. If necessary, researchers may need to debrief participants to explain the purpose of requesting particular demographic information.

### Location of Demographic Items

Often demographic information is gathered at the end of data collection, particularly when demographic questions could bias an individual's response. For example, asking participants to answer questions about income and ethnicity before they complete a survey may suggest to the participant that the researchers will explore responses on the basis of income and ethnicity.

*Marvin Lee and C. Melanie Schuele*

*See also* Dependent Variable; Independent Variable

**Further Readings**

Goldberg, W. A., Prause, J., Lucas-Thompson, R., & Himsel, A. (2008). Maternal employment and children's achievement in context: A meta-analysis of four decades of research. *Psychological Bulletin, 134,* 77–108.

Hart, D., Atkins, R., & Matsuba, M. K. (2008). The association of neighborhood poverty with personality change in childhood. *Journal of Personality & Social Psychology, 94,* 1048–1061.

## Dependent Variable

A dependent variable, also called an outcome variable, is the result of the action of one or more independent variables. It can also be defined as any outcome variable associated with some measure, such as a survey. Before providing an example, the relationship between the two (in an experimental setting) might be expressed as follows:

$$\mathrm{DV} = f(\mathrm{IV}_1 + \mathrm{IV}_2 + \mathrm{IV}_3 + \cdots + \mathrm{IV}_k),$$

where DV = the value of the dependent variable, $f$ = function of, and $\mathrm{IV}_k$ = the value of one or more independent variables.

In other words, the value of a dependent variable is a function of changes in one or more independent variables. The following abstract provides an example of a dependent variable and its interaction with a independent variable (ethnic origin):

> Ethnic origin is one factor that may influence the rate or sequence of infant motor development, interpretation of screening test results, and decisions regarding early intervention. The primary purpose of this study is to compare motor development screening test scores from infants of Asian and European ethnic origins. Using a cross-sectional design, the authors analyzed Harris Infant Neuromotor Test (HINT) scores of 335 infants of Asian and European origins. Factorial ANOVA results indicated no significant differences in test scores between infants from

these two groups. Although several limitations should be considered, results of this study indicate that practitioners can be relatively confident in using the HINT to screen infants of both origins for developmental delays. [Mayson, T. A., Backman, C. L., Harris, S. & Hayes, V. E. (2009). Motor development in Canadian infants of Asian and European ethnic origins. *Journal of Early Intervention, 31*(3), 199–214.]

In this study, the dependent variable is motor development as measured by the Harris Infant Neuromotor Test (HINT), and the independent variables are ethnic origin (with the two categorical levels of Asian origin and European origin). In this quasi-experimental study (since participants are preassigned), scores on the HINT are a function of ethnic origin.

In the following example, the dependent variable is a score on a survey reflecting how well survey participants believe that their students are prepared for professional work. Additional analyses looked at group differences in program length, but the outcome survey values illustrate what is meant in this context as a dependent variable.

This article presents results from a survey of faculty members from 2- and 4-year higher education programs in nine states that prepare teachers to work with preschool children. The purpose of the study was to determine how professors address content related to social-emotional development and challenging behaviors, how well prepared they believe graduates are to address these issues, and resources that might be useful to better prepare graduates to work with children with challenging behavior. Of the 225 surveys that were mailed, 70% were returned. Faculty members reported their graduates were prepared on topics such as working with families, preventive practices, and supporting social emotional development but less prepared to work with children with challenging behaviors. Survey findings are discussed related to differences between 2- and 4-year programs and between programs with and without a special education component. Implications for personnel preparation and future research are discussed. [Hemmeter, M. L., Milagros Santos, R. M., & Ostrosky, M. M. (2008). Preparing early childhood educators to address young children's social-emotional development and challenging behavior. *Journal of Early Intervention, 30*(4), 321–340.]

*Neil J. Salkind*

***See also*** Control Variables; Dichotomous Variable; Independent Variable; Meta-Analysis; Nuisance Variable; Random Variable; Research Hypothesis

**Further Readings**

Luft, H. S. (2004). Focusing on the dependent variable: Comments on "Opportunities and Challenges for Measuring Cost, Quality, and Clinical Effectiveness in Health Care," by Paul A. Fishman, Mark C. Hornbrook, Richard T. Meenan, and Michael J. Goodman. *Medical Care Research and Review, 61,* 144S–150S.

Sechrest, L. (1982). Program evaluation: The independent and dependent variables. *Counseling Psychologist, 10,* 73–74.

# Descriptive Discriminant Analysis

*Discriminant analysis* comprises two approaches to analyzing group data: descriptive discriminant analysis (DDA) and predictive discriminant analysis (PDA). Both use continuous (or intervally scaled) data to analyze the characteristics of group membership. However, PDA uses this continuous data to predict group membership (i.e., How accurately can a classification rule classify the current sample into groups?), while DDA attempts to discover what continuous variables contribute to the separation of groups (i.e., Which of these variables contribute to group differences and by how much?). In addition to the primary goal of discriminating among groups, DDA can examine the most parsimonious way to discriminate between groups, investigate the amount of variance accounted for by the discriminant variables, and evaluate the relative contribution of each discriminant (continuous) variable in classifying the groups.

For example, a psychologist may be interested in which psychological variables are most responsible for men's and women's progress in therapy.

For this purpose, the psychologist could collect data on therapeutic alliance, resistance, transference, and cognitive distortion in a group of 50 men and 50 women who report progressing well in therapy. DDA can be useful in understanding which variables of the four (therapeutic alliance, resistance, transference, and cognitive distortion) contribute to the differentiation of the two groups (men and women). For instance, men may be low on therapeutic alliance and high on resistance. On the other hand, women may be high on therapeutic alliance and low on transference. In this example, the other variable of cognitive distortion may not be shown to be relevant to group differentiation at all because it does not capture much difference among the groups. In other words, cognitive distortion is unrelated to how men and women progress in therapy. This is just a brief example of the utility of DDA in differentiating among groups.

DDA is a multivariate technique with goals similar to those of multivariate analysis of variance (MANOVA) and computationally identical to MANOVA. As such, all assumptions of MANOVA apply to the procedure of DDA. However, MANOVA can determine only *whether* groups are different, not *how* they are different. In order to determine how groups differ using MANOVA, researchers typically follow the MANOVA procedure with a series of analyses of variance (ANOVAs). This is problematic because ANOVAs are univariate tests. As such, several ANOVAs may need to be conducted, increasing the researcher's likelihood of committing Type I error (likelihood of finding a statistically significant result that is not really there). What's more, what makes multivariate statistics more desirable in social science research is the inherent assumption that human behavior has multiple causes and effects that exist simultaneously. Conducting a series of univariate ANOVAs strips away the richness that multivariate analysis reveals because ANOVA analyzes data as if differences among groups occur in a vacuum, with no interaction among variables. Consider the earlier example. A series of ANOVAs would assume that as men and women progress through therapy, there is no potential shared variance between the variables therapeutic alliance, resistance, transference, and cognitive distortion. And while MANOVA does account for this shared variance, it cannot tell the researcher how or where the differences come from.

This entry first describes discriminant functions and their statistical significance. Next, it explains the assumptions that need to be met for DDA. Finally, it discusses the computation and interpretation of DDA.

## Discriminant Functions

A discriminant function (also called a *canonical discriminant function*) is a weighted linear combination of discriminant variables, which can be written as

$$D = a + b_1x_1 + b_2x_2 + \cdots + b_nx_n + c, \qquad (1)$$

where $D$ is the discriminant score, $a$ is the intercept, the $b$s are the discriminant coefficients, the $x$s are discriminant variables, and $c$ is a constant. The discriminant coefficients are similar to beta weights in multiple regression and maximize the distance across the means of the grouping variable. The number of discriminant functions in DDA is $k - 1$, where $k$ is the number of groups or categories in the grouping variable, or the number of discriminant variables, whichever is less. For example, in the example of men's and women's treatment progress, the number of discriminant functions will be one because there are two groups and four discriminant variables, that is, min (1, 4), 1 is less than 4. In DDA, discriminant variables are optimally combined so that the first discriminant function provides the best discrimination across groups, the second function second best, and so on until all possible dimensions are assessed. These functions are orthogonal or independent from one another so that there will be no shared variance among them (i.e., no overlap of contribution to differentiation of groups). The first discriminant function will represent the most prevailing discriminating dimension, and later functions may also denote other important dimensions of discrimination.

The statistical significance of each discriminant function should be tested prior to a further evaluation of the function. Wilks's lambda is used to examine the statistical significance of functions. Wilks's lambda varies from 0 through 1, with 1 denoting the groups that have the same mean

discriminant function scores and 0 denoting those that have different mean scores. In other words, the smaller the value of Wilks's lambda, the more likely it is statistically significant and the better it differentiates between the groups. Wilks's lambda is the ratio of within-group variance to the total variance on the discrimiant variables and indicates the proportion of variance in the total variance that is not accounted for by differences of groups. A small lambda indicates the groups are well discriminated. In addition, 1 − Wilks's lambda is used as a measure of effect size to assess the practical significance of discriminant functions as well as the statistical significance.

## Assumptions

DDA requires seven assumptions to be met. First, DDA requires two or more mutually exclusive groups, which are formed by the grouping variable with each case belonging to only one group. It is best practice for groups to be truly categorical in nature. For example, sex, ethnic group, and state where someone resides are all categorical. Sometimes researchers force groups out of otherwise continuous data. For example, people aged 15 to 20, or income between $15,000 and $20,000. However, whenever possible, preserving continuous data where it exists and using categorical data as grouping variables in DDA is best. The second assumption states there must be at least two cases for each group.

The other five assumptions are related to discriminant variables in discriminant functions, as explained in the previous section. The third assumption states that any number of discriminant variables can be included in DDA as long as the number of discriminant variables is less than the sample size of the smallest group. However, it is generally recommended that the sample size be between 10 and 20 times the number of discriminant variables. If the sample is too small, the reliability of a DDA will be lower than desired. On the other hand, if the sample size is too large, statistical tests will turn out significant even for small differences. Fourth, the discriminant variables should be interval, or at least ordinal. Fifth, the discriminant variables are not completely redundant or highly correlated with each other. This assumption is identical to the absence of perfect multicollinearity assumption in multiple regression. If a discriminant variable is very highly correlated with another discriminant variable (e.g., $r > .90$), the variance–covariance matrix of the discriminant variables cannot be inverted. Then, the matrix is called *ill-conditioned*. Sixth, discriminant variables must follow the multivariate normal distribution, meaning that a discriminant variable should be normally distributed about fixed values of all the other discriminant variables. K. V. Mardia has provided measures of multivariate skewness and kurtosis, which can be computed to assess whether the combined distribution of discriminant variables is multivariate. Also, multivariate normality can be graphically evaluated. Seventh, DDA assumes that the variance–covariance matrices of discriminant variables are homogeneous across groups. This assumption intends to make sure that the compared groups are from the same population. If this assumption is met, any differences in a DDA analysis can be attributed to discriminant variables, but not to the compared groups. This assumption is analogous to the homogeneity of variance assumption in ANOVA. The multivariate Box's $M$ test can be used to determine whether the data satisfies this assumption. Box's $M$ test examines the null hypothesis that the variance–covariance matrices are not different across the groups compared. If the test is significant (e.g., the $p$ value is lower than .05), the null hypothesis can be rejected, indicating that the matrices are different across the groups. However, it is known that Box's $M$ test is very sensitive to even small differences in variance–covariance matrices when the sample size is large. Also, because it is known that DDA is robust against violation of this assumption, the $p$ value typically is set at a much lower level, such as .001. Furthermore, it is recognized that DDA is robust with regard to violation of the assumption of multivariate normality.

When data do not satisfy some of the assumptions of DDA, logistic regression can be used as an alternative. Logistic regression can answer the same kind of questions DDA answers. Also, it is a very flexible method in that it can handle both categorical and interval variables as discriminant variables and data under analysis do not need to meet assumptions of multivariate normality and equal variance–covariance matrices. It is also robust to unequal group size.

## Computation

When a DDA is conducted, a canonical correlation analysis is performed computationally that will determine discriminant functions and will calculate their associated eigenvalues and canonical correlations. Each discriminant function has its own eigenvalue. Eigenvalues, also called *canonical roots* or *characteristic roots*, denote the proportion of between-group variance explained by the respective discriminant functions, and their proportions add up to 100% for all discriminant functions. Thus, the ratio of two eigenvalues shows the relative differentiating power of their associated discriminant functions. For example, if the ratio is 1.5, then the discriminant function with the larger eigenvalue explains 50% more of the between-group variance in the grouping variable than the function with the smaller eigenvalue does. The canonical correlation is a correlation between the grouping variable and the discriminant scores that are measured by the composite of discriminant variables in the discriminant function. A high canonical correlation is associated with a function that differentiates groups well.

## Interpretation

Descriptive discriminant functions are interpreted by evaluating the standardized discriminant coefficients, the structure coefficients, and the centroids. Standardized discriminant coefficients represent weights given to each discriminant variable in proportion to how well it differentiates groups and to how many groups it differentiates. Thus, more weight will be given to a discriminant variable that differentiates groups better. Because standardized discriminant coefficients are semipartial coefficients as standardized beta coefficients in multiple regression and expressed as $z$ scores with a mean of zero and a standard deviation of 1, they represent the relative significance of each discriminant variable to its discriminant function. The greater the coefficient, the larger is the contribution of its associated discriminant variable to the group differentiation. However, these standardized beta coefficients do not tell us the absolute contribution of each variable to the discriminant function. This becomes a serious problem when any two discriminant variables in the discriminant function are highly correlated and thus have a large amount of shared variance between them.

Because of these issues, it is recommended that researchers consider the structure coefficients as well as the standardized discriminant coefficients to determine which variables define the nature of a specific discriminant function. The structure coefficients, or the factor structure coefficients, are not semipartial coefficients like the standardized discriminant coefficients but are whole coefficients, like correlation coefficients. The structure coefficients represent uncontrolled association between the discriminant functions and the discriminant variables. Because the factor coefficients are correlations between the discriminant variables and the discriminant functions, they can be conceptualized as factor loadings on latent dimensions, as in factor analysis.

However, in some cases these two coefficients do not agree. For example, the standardized discriminant coefficient might tell us that a specific discriminant variable differentiates groups most, but the structure coefficient might indicate the opposite. A body of previous research says the standardized discriminant coefficient and the structure coefficient can be unreliable with a small sample size, such as when the ratio of the number of subjects to the number of discriminant variables drops below 20:1. Besides increasing the sample size, Maurice Tatsuoka has suggested that the standardized discriminant coefficient be used to investigate the nature of each discriminant variable's contribution to group discrimination and that the structure coefficients be used to assign substantive labels to the discriminant functions.

Although these two types of coefficients inform us of the relationship between discriminant functions and their discriminant variables, they do not tell us which of the groups the discriminant functions differentiate most or least. In other words, the coefficients do not provide any information on the grouping variable. To return to the earlier example of treatment progress, the DDA results demonstrated that therapeutic alliance, resistance, and transference are mainly responsible for the differences between men and women in the discriminant scores. However, we still need to explore which group has more or less of these three psychological traits that are found to be differentiating. Thus, we need to investigate the mean discriminant function scores for each group, which are called *group*

*centroids*. The nature of the discrimination for each discriminant function can be examined by looking at different locations of centroids. For example, a certain group that has the highest and lowest values of centroids on a discriminant function will be best discriminated on that function.

*Seong-Hyeon Kim and Alissa Sherry*

***See also*** Multivariate Analysis of Variance (MANOVA)

**Further Readings**

Brown, M. T., & Wicker, L. R. (2000). Discriminant analysis. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of multivariate statistics and mathematical modeling* (pp. 209–235). San Diego, CA: Academic Press.

Huberty, C. J. (1994). *Applied discriminant analysis*. New York: Wiley.

Sherry, A. (2006). Discriminant analysis in counseling psychology research. *The Counseling Psychologist, 5,* 661–683.

# Descriptive Statistics

Descriptive statistics are commonly encountered, relatively simple, and for the most part easily understood. Most of the statistics encountered in daily life, in newspapers and magazines, in television, radio, and Internet news reports, and so forth, are descriptive in nature rather than inferential. Compared with the logic of inferential statistics, most descriptive statistics are somewhat intuitive. Typically the first five or six chapters of an introductory statistics text consist of descriptive statistics (means, medians, variances, standard deviations, correlation coefficients, etc.), followed in the later chapters by the more complex rationale and methods for statistical inference (probability theory, sampling theory, $t$ and $z$ tests, analysis of variance, etc.)

Descriptive statistical methods are also foundational in the sense that inferential methods are conceptually dependent on them and use them as their building blocks. One must, for example, understand the concept of variance before learning how analysis of variance or $t$ tests are used for statistical inference. One must understand the descriptive correlation coefficient before learning how to use regression or multiple regression inferentially. Descriptive statistics are also complementary to inferential ones in analytical practice. Even when the analysis draws its main conclusions from an inferential analysis, descriptive statistics are usually presented as supporting information to give the reader an overall sense of the direction and meaning of significant results.

Although most of the descriptive building blocks of statistics are relatively simple, some descriptive methods are high level and complex. Consider multivariate descriptive methods, that is, statistical methods involving multiple dependent variables, such as factor analysis, principal components analysis, cluster analysis, canonical correlation, or discriminant analysis. Although each represents a fairly high level of quantitative sophistication, each is primarily descriptive. In the hands of a skilled analyst, each can provide invaluable information about the holistic patterns in data. For the most part, each of these high-level multivariate descriptive statistical methods can be matched to a corresponding inferential multivariate statistical method to provide both a description of the data from a sample and inferences to the population; however, only the descriptive methods are discussed here.

The topic of descriptive statistics is therefore a very broad one, ranging from the simple first concepts in statistics to the higher reaches of data structure explored through complex multivariate methods. The topic also includes graphical data presentation, exploratory data analysis (EDA) methods, effect size computations and meta-analysis methods, esoteric models in mathematical psychology that are highly useful in basic science experimental psychology areas (such as psychophysics), and high-level multivariate graphical data exploration methods.

## Graphics and EDA

Graphics are among the most powerful types of descriptive statistical devices and often appear as complementary presentations even in primarily inferential data analyses. Graphics are also highly useful in the exploratory phase of research, forming an essential part of the approach known as EDA.

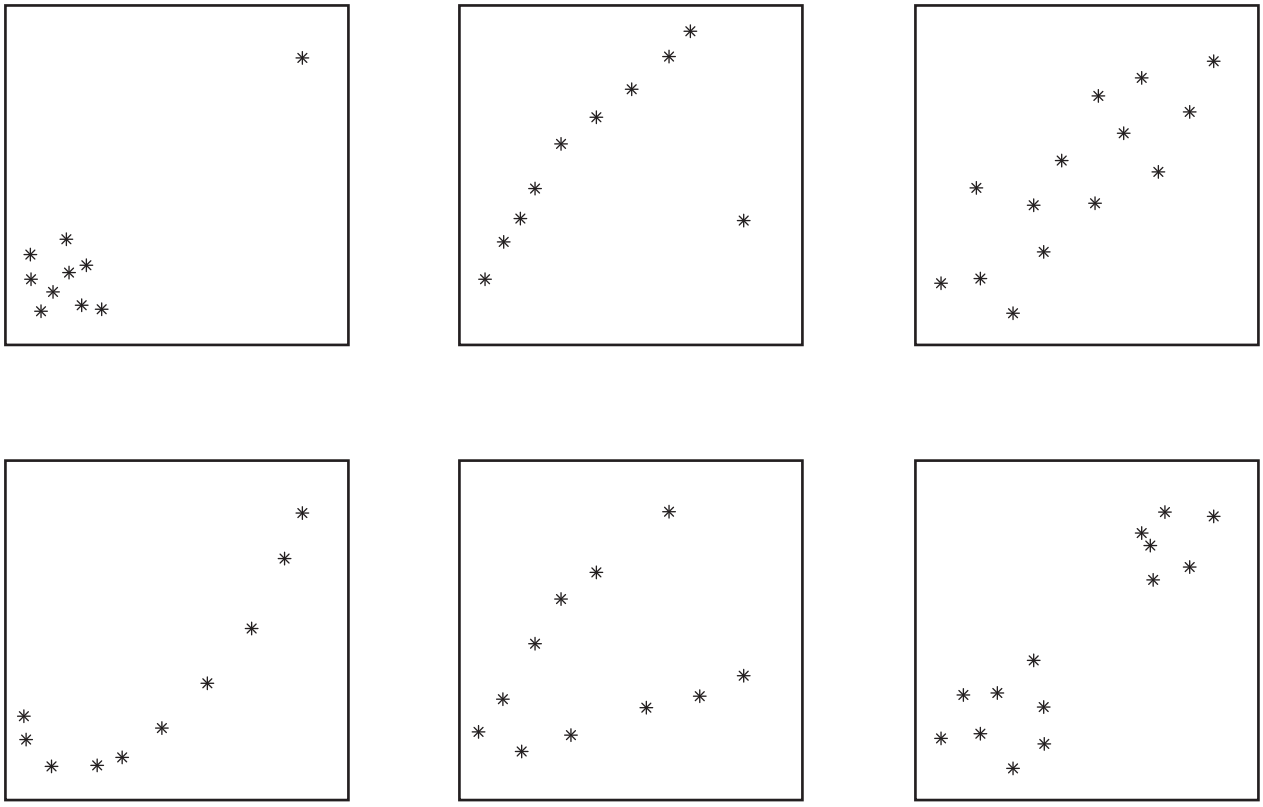


**Figure 1** Scatterplots of Six Different Bivariate Data Configurations That All Have the Same Pearson Product-Moment Correlation Coefficient

### *Graphics*

Many of the statistics encountered in everyday life are visual in form—charts and graphs. Descriptive data come to life and become much clearer and substantially more informative through a well-chosen graphic. One only has to look through a column of values of the Dow Jones Industrial Average closing price for the past 30 days and then compare it with a simple line graph of the same data to be convinced of the clarifying power of graphics. Consider also how much more informative a scatterplot is than the correlation coefficient as a description of the bivariate relationship between two variables. Figure 1 uses bivariate scatterplots to display six different data sets that all have the same correlation coefficient. Obviously there is much more to be known about the structural properties of a bivariate relationship than merely its strength (correlation), and much of this is revealed in a scatterplot.

Graphics have marvelous power to clarify, but they can also be used to obfuscate. They can be highly misleading and even deceitful, either intentionally or unintentionally. Indeed, Darrell Huff's classic book *How to Lie With Statistics* makes much of its case by demonstrating deceitful graphing practices. One of them is shown in Figure 2. Suppose that a candidate for election to the office of sheriff were to show the failings of the incumbent with a graph like the one on the left side of Figure 2. At first look, it appears that crime has increased at an alarming rate during the incumbent's tenure, from 2007 to 2009. The 2009 bar is nearly 3 times as high as the bar for 2007. However, with a more careful look, it is apparent that we have in fact magnified a small segment of the y-axis by restricting the range (from 264 crimes per 100,000 people to 276). When the full range of y-axis values is included, together with the context of the surrounding years, it becomes apparent that what appeared

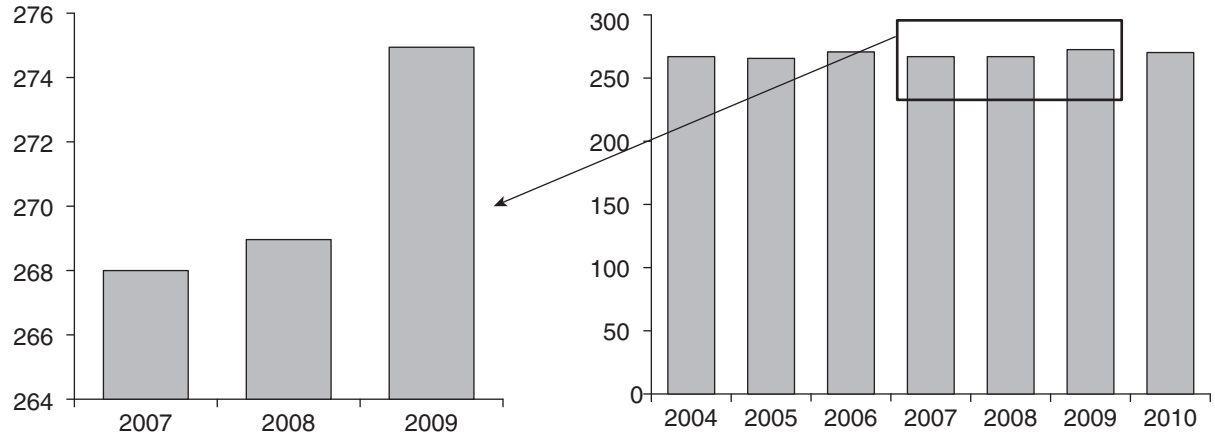



**Figure 2** Two Bar Graphs Demonstrating the Importance of Proper Scaling
*Note:* On right, crimes per 100,000 persons from a 7-year-series bar graph; on left, same data clipped to 3 years, with the *y*-axis trimmed to a misleadingly small magnitude range to create an incorrect impression.

to be a strongly negative trend is more reasonably attributed to random fluctuation.

Although the distortion just described is intentional, similar distortions are common through oversight. In fact, if one enters the numerical values from Figure 2 into a spreadsheet program and creates a bar graph, the *default graph* employs the restricted range shown in the left-hand figure. It requires a special effort to present the data accurately. Graphics can be highly illuminating, but the caveat is that one must use care to ensure that they are not misleading. The popularity of Huff's book indicates that he has hit a nerve in questioning the veracity in much of statistical presentation.

The work of Edward Tufte is also well known in the statistical community, primarily for his compelling and impressive examples of best practices in the visual display of quantitative information. Although he is best known for his examples of good graphics, he is also adept in identifying a number of the worst practices, such as what he calls "chartjunk," or the misleading use of rectangular areas in picture charts, and the often mindless use of PowerPoint in academic presentations.

### *EDA*

Graphics have formed the basis of one of the major statistical developments of the past 50 years: EDA. John Tukey is responsible for much of this development, with his highly creative graphical methods, such as the stem-and-leaf plot and the box-and-whisker plot, as shown on the left and right, respectively, in Figure 3. The stem-and-leaf (with the 10s-digit stems on the left of the line and the units-digit leaves on the right) has the advantage of being both a table and a graph. The overall shape of the stem-and-leaf plot in Figure 3 shows the positive skew in the distribution, while the precise value of each data point is preserved by numerical entries. The box-and-whisker plots similarly show the overall shape of a distribution (the two in Figure 3 having opposite skew) while identifying the summary descriptive statistics with great precision. The box-and-whisker plot can also be effectively combined with other graphs (such as attached to the *x*- and the *y*-axes of a bivariate scatterplot) to provide a high level of convergent information. These methods and a whole host of other illuminating graphical displays (such as run charts, Pareto charts, histograms, MultiVari charts, and many varieties of scatterplots) have become the major tools of data exploration.

Tukey suggested that data be considered decomposable into rough and smooth elements (data = rough + smooth). In a bivariate relationship, for example, the regression line could be considered the smooth component, and the deviations from regression the rough

component. Obviously, a description of the smooth component is of value, but one can also learn much from a graphical presentation of the rough.

Tukey contrasted EDA with confirmatory data analysis (the testing of hypotheses) and saw each as having its place, much like descriptive and inferential statistics. He referred to EDA as a reliance on display and an attitude.

### *The Power of Graphicity*

Although graphical presentations can easily go astray, they have much potential explanatory power and exploratory power, and some of the best of the available descriptive quantitative tools are in fact graphical in nature. Indeed, it has been persuasively argued, and some evidence has been given, that the use of graphs in publications both within psychology and across other disciplines correlates highly with the "hardness" of those scientific fields. Conversely, an inverse relation is found between hardness of subareas of psychology and the use of inferential statistics and data tables, indicating that the positive correlation of graphicity with hardness is not due to quantification and that perhaps inferential methods are often used in an attempt to deal with inadequate data.

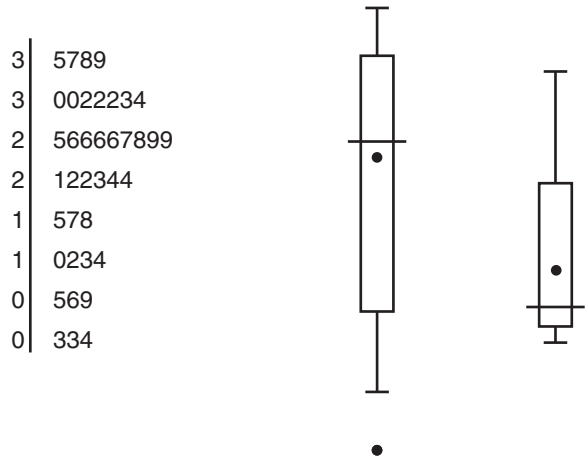



**Figure 3** Stem-and-Leaf Plot (Left); Two Box-and-Whisker Plots (Right)

*Note:* The box-and-whisker plots consist of a rectangle extending from the 1st quartile to the 3rd quartile, a line across the rectangle (at the 2nd quartile, or median), and "whisker" lines at each end marking the minimum and maximum values.

The available selection of graphical descriptive statistical tools is obviously broad and varied. It includes simple graphical inscription devices—such things as bar graphs, line graphs, histograms, scatterplots, box-and-whisker plots, and stem-and-leaf plots, as just discussed—and also high-level ones. Over the past century a number of highly sophisticated multidimensional graphical methods have been devised. These include principal components plots, multidimensional scaling plots, cluster analysis dendrograms, Chernoff faces, Andrews plots, time series profile plots, and generalized draftsman's displays (also called multiple scatterplots), to name a few.

Hans Rosling, a physician with broad interests, has created a convincing demonstration of the immense explanatory power of so simple a graph as a scatterplot, using it to tell the story of economic prosperity and health in the development of the nations of the world over the past two centuries. His lively narration of the presentations accounts for some of their impact, but such data stories can be clearly told with, for example, a time-series scatterplot of *balloons* (the diameter of each representing the population size of a particular nation) floating in a bivariate space of fertility rate ($x$-axis) and life expectancy ($y$-axis). The time-series transformations of this picture play like a movie, with labels of successive years ("1962," "1963," etc.) flashing in the background.

## Effect Size, Meta-Analysis, and Accumulative Data Description

Effect size statistics are essentially descriptive in nature. They have evolved in response to a logical gap in established inferential statistical methods. Many have observed that the alternative hypothesis is virtually always supported if the sample size is large enough and that many published and statistically significant results do not necessarily represent strong relationships. To correct this somewhat misleading practice, William L. Hays, in his 1963 textbook, introduced methods for calculating effect size.

In the 30 years that followed, Jacob Cohen took the lead in developing procedures for effect size estimation and power analysis. His work in turn

led to the development of meta-analysis as an important area of research—comparisons of the effect sizes from many studies, both to properly estimate a summary effect size value and to assess and correct bias in accumulated work. That is, even though the effect size statistic itself is descriptive, inferential data-combining methods have been developed to estimate effect sizes on a population level.

Another aspect of this development is that the recommendation that effect sizes be reported has begun to take on a kind of ethical force in contemporary psychology. In 1996, the American Psychological Association Board of Scientific Affairs appointed a task force on statistical inference. Its report recommended including effect size when reporting a $p$ value, noting that reporting and analyzing effect size is imperative to good research.

## Multivariate Statistics and Graphics

Many of the commonly used multivariate statistical methods, such as principal components analysis, some types of factor analysis, cluster analysis, multidimensional scaling, discriminant analysis, and canonical correlation, are essentially descriptive in nature. Each is conceptually complex, useful in a practical sense, and mathematically interesting. They provide clear examples of the farther reaches of sophistication within the realm of descriptive statistics.

### *Principal Components Analysis and Factor Analysis*

Factor analysis has developed within the discipline of psychology over the past century in close concert with psychometrics and the mental testing movement, and it continues to be central to psychometric methodology. Factor analysis is in fact not one method but a family of methods (including principal components) that share a common core. The various methods range from entirely descriptive (principal components, and also factor analysis by the principal components method) to inferential (common factors method, and also maximum likelihood method). Factors are extracted by the

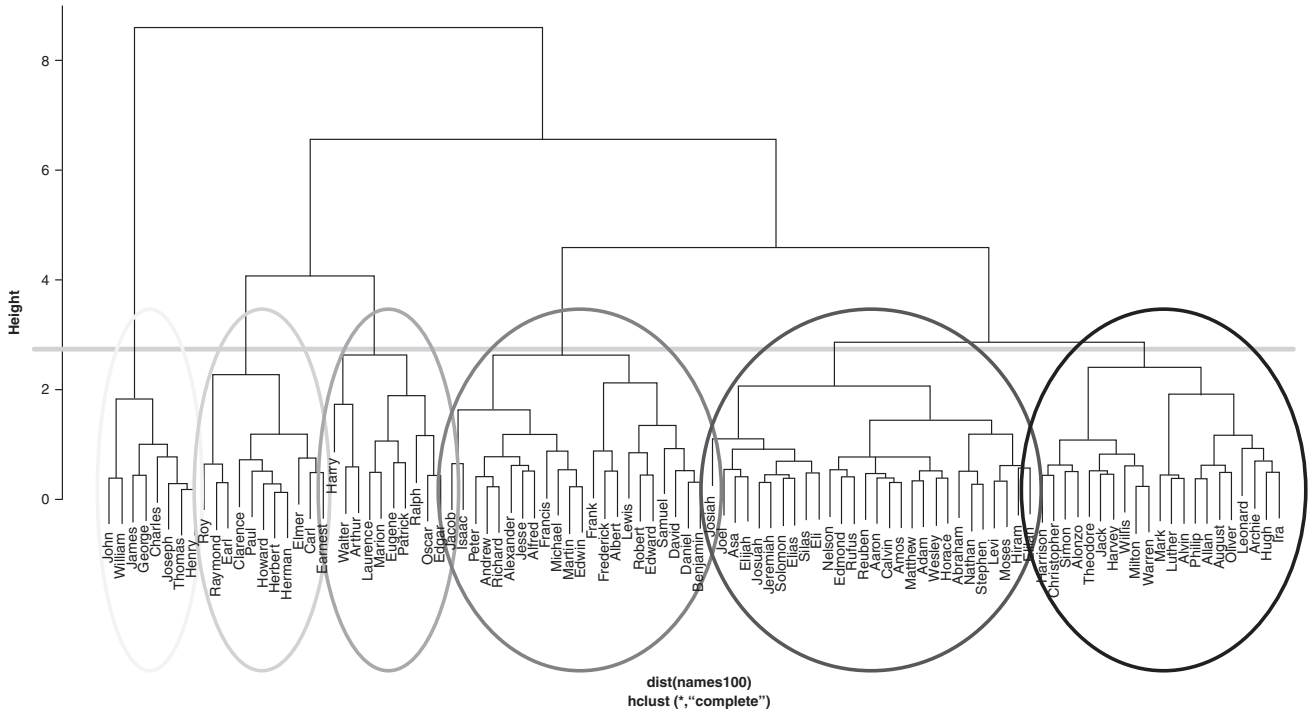



**Figure 4** Cluster Analysis Dendrogram of the Log-Frequencies of the 100 Most Frequent Male Names in the United States in the 19th Century

maximum likelihood method to account for as much variance as possible in *the population correlation matrix*. Principal components and the principal components method of factor analysis are most often employed, for descriptive ends, in creating multivariate graphics.

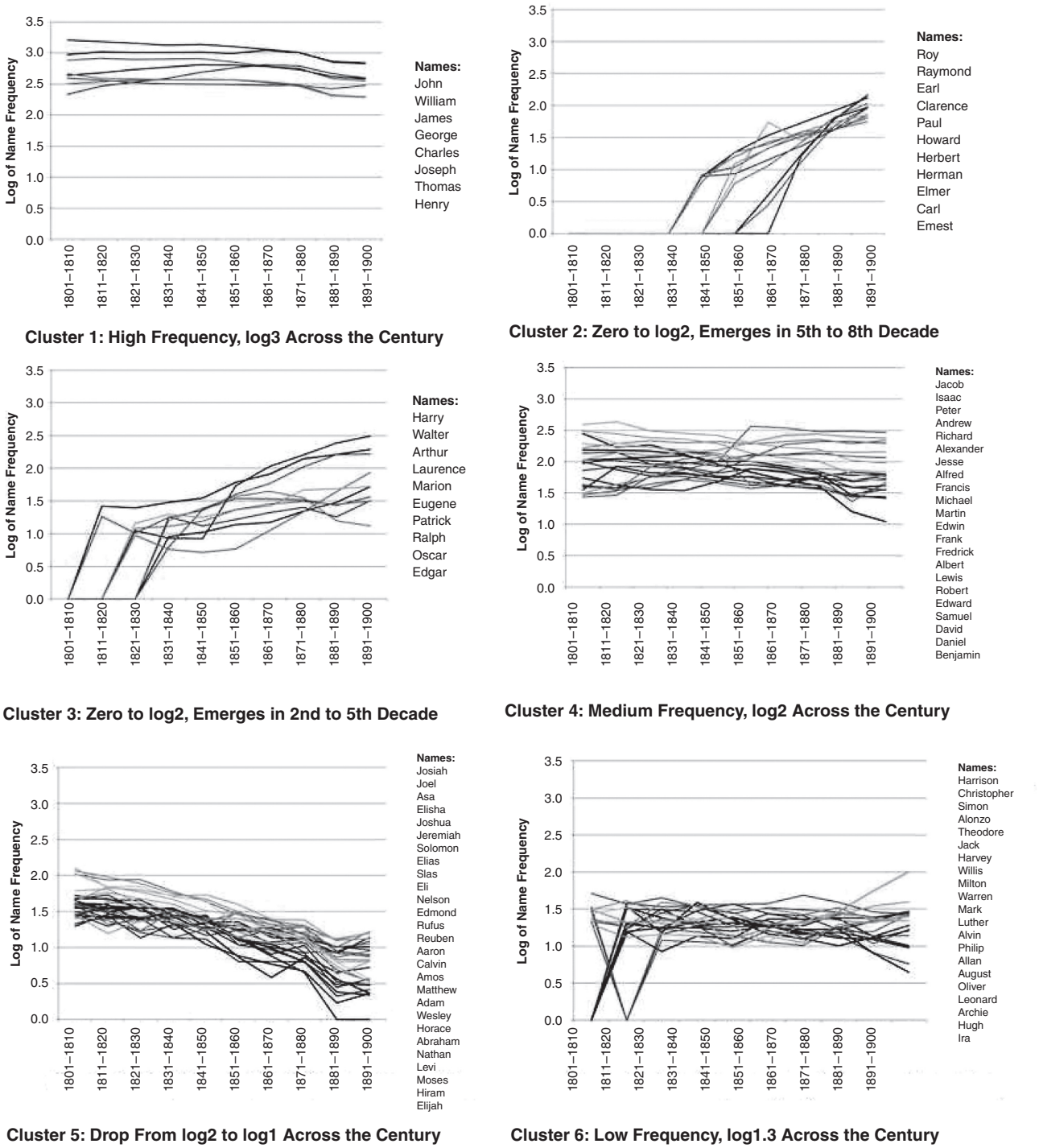



**Figure 5** Semantic Space of Male Names Defined by the Vectors for the 10 Decades of the 19th Century

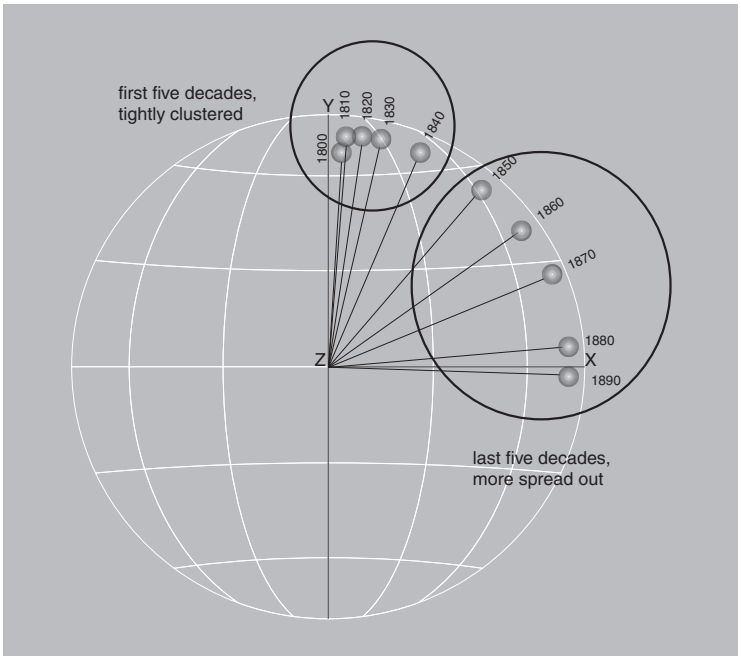



**Figure 6** Location of the 100 Most Frequent 19th Century Male Names Within the Semantic Space of 10 Decades

### *Other Multivariate Methods*

A number of other multivariate methods are also primarily descriptive in their focus and can be effectively used to create multivariate graphics, as several examples will illustrate. *Cluster analysis* is a method for finding natural groupings of objects within a multivariate space. It creates a graphical representation of its own, the *dendrogram,* but it can also be used to group points within a scatterplot. *Discriminant analysis* can be used graphically in essentially the same way as factor analysis and principal components, except that the factors are derived to maximally separate known groups rather than to maximize variance. *Canonical correlation* can be thought of as a double factor analysis in which the factors from an *X* set of variables are calculated to maximize their correlation with corresponding factors from a *Y* set of variables. As such, it can form the basis for multivariate graphical devices for comparing entire sets of variables.

### *Multivariate Graphics*

A simple example, a $100 \times 10$ matrix of name frequencies, illustrates several multivariate graphs. This matrix, taken from U.S. Census data from the 19th century, contains frequencies per 10,000 names for each decade (columns) for each of the top 100 male names (rows). Figure 4 is a cluster analysis dendrogram of these names, revealing six clusters.

Each of the six clusters is shown as a profile plot in Figure 5, with a collage of line plots tracing the trajectory of each name within the cluster. Clearly, the cluster analysis separates the groups well. Figure 6 is a vector plot from a factor analysis, in which two factors account for 93.3% of the variance in the name frequency pattern for ten decades. The vectors for the first three or four decades of the century are essentially vertical, with the remaining decades fanning out sequentially to the right and with the final decade flat horizontally to the right. A scatterplot (not shown here) of the 100 names within this same two-factor space reveals that the names group well within the six clusters, with virtually no overlap among clusters.

*Bruce L. Brown*

***See also*** Bar Chart; Box-and-Whisker Plot; Effect Size, Measures of; Exploratory Data Analysis; Exploratory Factor Analysis; Mean; Median; Meta-Analysis; Mode; Pearson Product-Moment Correlation Coefficient; Residual Plot; Scatterplot; Standard Deviation; Variance

**Further Readings**

Brown, B. L. (in press). *Multivariate analysis for the biobehavioral and social sciences*. New York: Wiley.

Cudeck, R., & MacCallum, R. C. (2007). Preface. In R. Cudeck & R. C. MacCallum (Eds.), *Factor analysis at 100: Historical developments and future directions*. Mahwah, NJ: Lawrence Erlbaum.

Huff, D. (1954). *How to lie with statistics*. New York: Norton.

Kline, P. (1993). *The handbook of psychological testing*. London: Routledge.

Smith, L. D., Best, I. A., Stubbs, A., Archibald, A. B., & Roberson-Nay, R. (2002). Constructing knowledge:

The role of graphs and tables in hard and soft psychology. *American Psychologist, 57*(10), 749–761.
Tufte, E. R. (2001). *The visual display of quantitative information* (2nd ed.). Cheshire, CT: Graphics Press.
Tukey, J. W. (1980). We need both exploratory and confirmatory. *American Statistician, 34*(1), 23–25.

# Dichotomous Variable

A dichotomous variable, a special case of categorical variable, consists of two categories. The particular values of a dichotomous variable have no numerical meaning. A dichotomous variable can either be naturally existing or constructed by a researcher through recoding a variable with more variation into two categories. A dichotomous variable may be either an independent or a dependent variable, depending on its role in the research design. The role of the dichotomous variable in the research design has implications for the selection of appropriate statistical analyses. This entry focuses on how a dichotomous variable may be defined or coded and then outlines the implications of its construction for data analysis.

## Identification, Labeling, and Conceptualization of a Dichotomous Variable

### *Identification and Labeling*

A dichotomous variable may also be referred to as a categorical variable, a nonmetric variable, a grouped dichotomous variable, a classification variable, a dummy variable, a binary variable, or an indicator variable. Within a data set, any coding system can be used that assigns two different values.

### *Natural Dichotomous Variables*

Natural dichotomous variables are based on the nature of the variable and can be independently determined. These variables tend to be nominal, discrete categories. Examples include whether a coin toss is heads or tails, whether a participant is male or female, or whether a participant did or did not receive treatment. Naturally dichotomous variables tend to align neatly with an inclusive criterion or condition and require limited checking of the data for reliability.

### *Constructed Dichotomous Variables*

Dichotomous variables may be constructed on the basis of conceptual rationalizations regarding the variables themselves or on the basis of the distribution of the variables in a particular study.

#### Construction Based on Conceptualization

While studies may probe the frequency of incidence of particular life experiences, researchers may provide a conceptually or statistically embedded rationale to support reducing the variation in the range of the distribution into two groups. The conceptual rationale may be rooted in the argument that there is a qualitative difference between participants who did or did not receive a diagnosis of depression, report discrimination, or win the lottery. The nature of the variable under study may suggest the need for further exploration related to the frequency with which participants experienced the event being studied, such as a recurrence of depression, the frequency of reported discrimination, or multiple lottery winnings. Nonetheless, the dichotomous variable allows one to distinguish qualitatively between groups, with the issue of the multiple incidence or frequency of the reported event to be explored separately or subsequently.

#### Construction Based on Distribution

The original range of the variable may extend beyond a binomial distribution (e.g., frequency being recorded as an interval such as never, sometimes, often, or with an even broader range when a continuous variable with possible values of 1–7 may be reduced to two groups of 1–4 and 5–7). An analysis of the standard deviation and shape of the frequency distribution (i.e., as represented through a histogram, box-plot, or stem-and-leaf diagram) may suggest that it would be useful to recode the variable into two values. This recoding may take several forms, such as a simple median split (with 50% of scores receiving one value and the other 50% receiving the other value), or other divisions based on the distribution of the data (e.g., 75% vs. 25% or 90% vs. 10%) or other conceptual reasons. For example, single or low-frequency events (e.g., adverse effects of a treatment) may be contrasted with high-frequency

events. The recoding of a variable with a range of values into a dichotomous variable may be done intentionally for a particular analysis, with the original values and range of the variable maintained in the data set for further analysis.

## Implications for Statistical Analysis

The role of the dichotomous variable within the research design (i.e., as an independent or dependent variable), as well as the nature of the sample distribution (i.e., normally or nonlinearly distributed), influences the type of statistical analyses that should be used.

### *Dichotomous Variables as Independent Variables*

In the prototypical experimental or quasi-experimental design, the dependent variable represents behavior that researchers measure. Depending on the research design, a variety of statistical procedures (e.g., correlation, linear regression, and analyses of variance) can explore the relationship between a particular dependent variable (e.g., school achievement) and a dichotomous variable (e.g., the participant's sex or participation in a particular enrichment program). How the dichotomous variable is accounted for (e.g., controlled for, blocked) will be dictated by the particular type of analysis implemented.

### *Dichotomous Variables as Dependent Variables*

When a dichotomous variable serves as a dependent variable, there is relatively less variation in the predicted variable, and consequently the data do not meet the requirements of a normal distribution and linear relationship between the variables. When this occurs, there may be implications for the data analyses selected. For instance, if one were to assess how several predictor variables relate to a dichotomous dependent variable (e.g., whether a behavior is observed), then procedures such as logistic regression should be used.

## Applications

Dichotomous variables are prominent features of a research design. They may either occur naturally (e.g., true/false or male/female) or may be assigned randomly by the researcher to address a range of research issues (e.g., sought treatment vs. did not seek treatment or sought treatment between zero and two times vs. sought treatment three or more times). How a dichotomous variable is conceptualized and constructed within a research design (i.e., as an independent or a dependent variable) will affect the type of analyses appropriate to interpret the variable and describe, explain, or predict based in part on the role of the dichotomous variable. Because of the arbitrary nature of the value assigned to the dichotomous variable, it is imperative to consult the descriptive statistics of a study in order to fully interpret findings.

*Mona M. Abo-Zena*

***See also*** Analysis of Variance (ANOVA); Correlation; Covariate; Logistic Regression; Multiple Regression

### Further Readings

Budesco, D. V. (1985). Analysis of dichotomous variables in the presence of serial dependence. *Psychological Bulletin, 73*(3), 547–561.

Meyers, L. S., Gamst, G., & Guarnino, A. J. (2006). *Applied multivariate research: Design and interpretation*. Thousand Oaks, CA: Sage.

# Differential Item Functioning

Item bias represents a threat to the validity of test scores in many different disciplines. An item is considered to be biased if the item unfairly favors one group over another. More specifically, an item is considered to be biased if two conditions are met. First, performance on the item is influenced by sources other than differences on the construct of interest that are deemed to be detrimental to one group. Second, this extraneous influence results in differential performance across identifiable subgroups of examinees.

The use of the term *bias* refers to various contexts, both statistical and social. From a statistical point of view, an item is said to be biased if the expected test or item scores are not the same for subjects from different subpopulations, given the

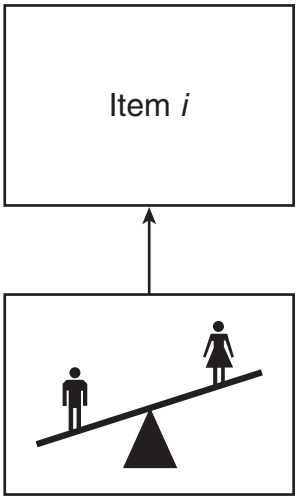



**Figure 1** An Illustration of Gender Effect

same level of trait on the instrument of interest. Thus, bias is not simply a difference between the means of item scores for subjects from different subpopulations. Group mean differences on an item could simply indicate differences in their ability on the construct the item is measuring. In order to show the presence of bias, one must show that groups continue to differ in their performance on an item or test even after their ability levels are controlled for. From a social point of view, an item is said to be biased if this difference is evaluated as being harmful to one group more than other groups.

In most psychometric research, there is an interest in detecting bias at the item level. One application of this would be in test development. Items that show bias can be reformulated or removed from the instrument. By considering bias at only the test level, one faces the real possibility of missing bias for a particular item. Furthermore, by considering bias on the item level, it is possible to see whether certain items are biased against certain subpopulations.

One characteristic of bias is differential item functioning (DIF), in which examinees from different groups have differing probabilities of success on an item after being matched on the ability of interest. DIF is a necessary but insufficient condition for item bias. If an item is biased, then DIF is present. However, the presence of DIF does not imply item bias in and of itself.

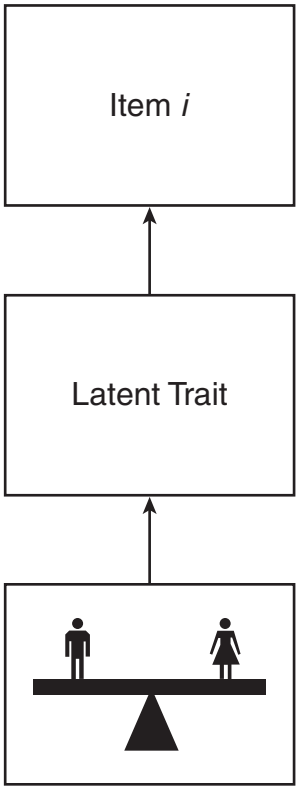



**Figure 2** An Illustration of No Gender Effect Controlling for Latent Trait

An illustration of DIF is given in Figures 1 through 3. In this example, suppose there are two groups of subjects (e.g., men and women) that have different probability of a dichotomous response on an item *i,* illustrated in Figure 1. A heavier weight signifies a higher probability of getting the item correct. In Figure 1, men have a higher probability of getting this particular item correct.

Because this item is an indicator of some latent, then the difference between the two groups is possibly attributable to the latent trait. Therefore, controlling for this latent trait (matching criterion) should remove the relationship between the gender and the item score. If this is the case, the item is measurement invariant across the groups. This is illustrated in Figure 2.

However, if the relationship between gender and the item remains after controlling for the latent trait, then DIF is present. That is, the item

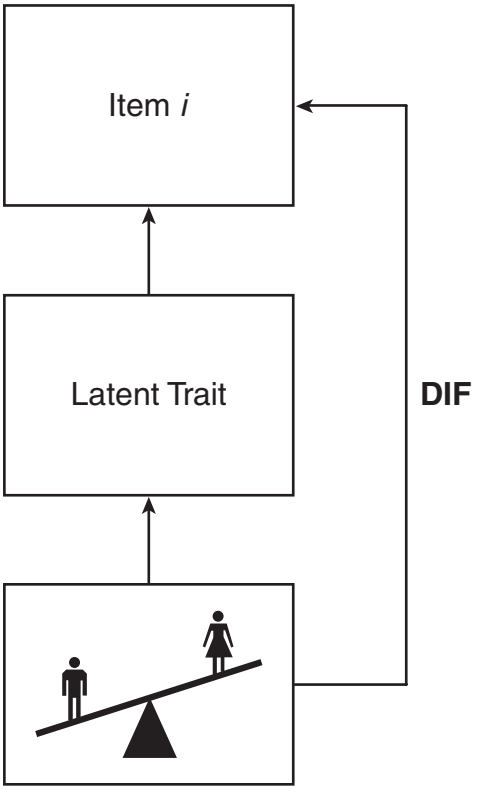



**Figure 3** An illustration of Differential Item Functioning

measures something in addition to the latent trait that is differentially related to the group variable. This is shown in Figure 3.

The remainder of this entry defines DIF and its related terminology, describes the use of DIF for polytomous outcomes, and discusses the assessment and measurement of DIF.

## Definition of DIF

DIF is one way to consider the different impact an item may have on various subpopulations. One could consider DIF as the statistical manifestation of bias, but not the social aspect. An item is said to show DIF when subjects from two subpopulations have different expected scores on the same item after controlling for ability. Using item response theory (IRT) terminology, if a non-DIF item has the same item response function between groups, then subjects having the same ability would have equal probability of getting the item correct. However, items that have different item response functions between groups indicate that DIF is present and that the items may be biased.

It is important to note that an item might show DIF yet not be biased. DIF is a necessary condition for bias, yet biased items reflect a more stringent interpretation of the severity of impact of the subpopulation difference. The presence of DIF is detected by statistical means, yet an item is considered biased only on interpretation of the meaning of the detected difference. An item with DIF might not always indicate bias. The item could simply indicate a multidimensional facet of the item or test. Thus, DIF analysis detects only possibly biased items. Akihito Kamata and Brandon K. Vaughn have suggested that an item with DIF should be referred to as a "possibly biased item," or simply as a "DIF item."

## Terminology

DIF analysis typically compares two groups: a focal group and a reference group. The focal group is defined as the main group of interest, whereas the reference group is a group used for comparison purposes. The statistical methodology of DIF assumes that one controls for the trait or ability levels between these two groups. Most research uses the term *ability level* for either ability levels or trait levels, even though in specific situations one term might be more precise. The ability level is used to match subjects from the two groups so that the effect of ability is controlled. Thus, by controlling for ability level, one may detect group differences that are not confounded by the ability. This ability level is aptly referred to as the matching criterion.

The matching criterion might be one of many different indices of interest, yet typically the total test performance or some estimate of trait levels (as in the case of attitudinal measures) is used. In some instances, an external measure might be used as the matching criterion if it can be shown that the measure is appropriate to account for the ability levels of the groups of interest. In addressing the issue of using test scores as the matching criterion, the matching

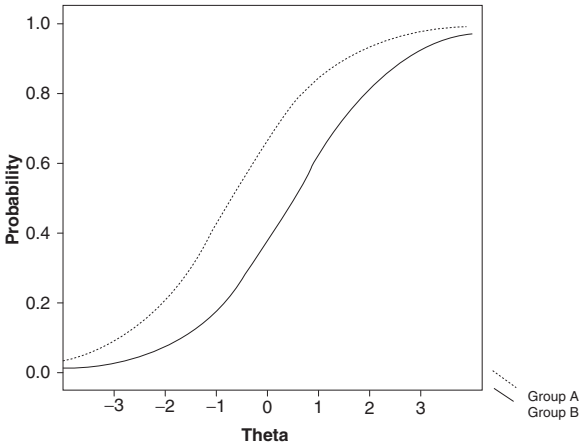



**Figure 4** Example of Uniform DIF

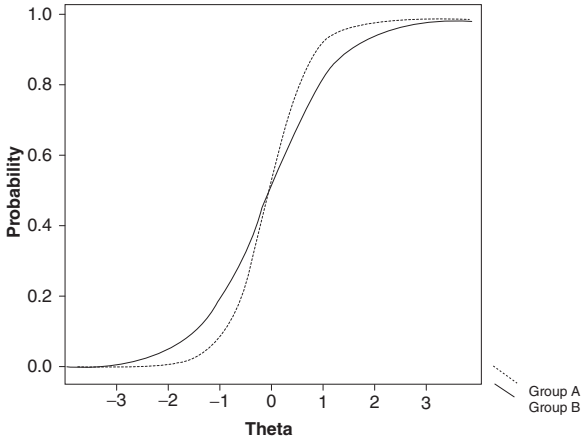



**Figure 5** Example of Nonuniform DIF

criterion should be free of DIF items. This can be problematic in the typical case in which the items undergoing DIF analysis are the very items that form the matching criterion. In such a situation, the matching criterion should undergo a "purification" process, in which a preliminary DIF analysis is performed to rid the matching criterion of any DIF items.

The phrase *uniform DIF* refers to a type of DIF in which the magnitude of group difference is the same across ability levels. Using IRT ideas, uniform DIF occurs when there is no interaction between group and item characteristic curves, as represented in Figure 4. In contrast, the phrase *nonuniform DIF* refers to a type of DIF in which the magnitude of the group difference is not consistent across ability levels. From an IRT perspective, nonuniform DIF would result in crossing item characteristic curves. This is illustrated in Figure 5. Nonuniform DIF can be thought of as an interaction effect between the group and the ability level.

## DIF for Polytomous Outcomes

Although traditional DIF procedures involve dichotomously scored items, DIF can also be considered for polytomously scored data (e.g., Likert-type scales). Polytomously scored data have the additional consideration that subjects can respond to or be labeled with more than two categories on a given item. For dichotomous data, the consideration of DIF is more simplistic as there are only two outcomes. But for polytomous outcomes, there is a possibility of an *inner-response DIF* (IDIF). That is, there is the possibility that DIF does not exist uniformly across all response categories but may exist for certain responses within that item. Figure 6 illustrates an example in which a particular 4-point Likert-type item displays DIF on lower ordinal responses but not on higher ordinal responses. This type of DIF can be referred to as a *lower IDIF*. This can exist, as an illustration, when the focal group tends to differentially vary in successfully scoring lower ordinal scores on an attitudinal measurement as compared to a reference group, while both groups have similar success in upper ordinal scoring categories.

Figure 7 illustrates a *balanced IDIF*, in which the nature of DIF changes for both extreme ordinal responses.

In this example, there is potential bias against women on the lower ordinal responses, and potential bias against men on the upper responses. Other types of IDIF patterns are possible. For example, *upper IDIF* would indicate potential bias on the upper ordinal responses, while *consistent IDIF* would indicate that the DIF effect is approximately the same for all ordinal responses. Patterns in IDIF are not always present, however. In some situations, IDIF may be present only between certain ordinal responses and not others, with no discernible pattern.

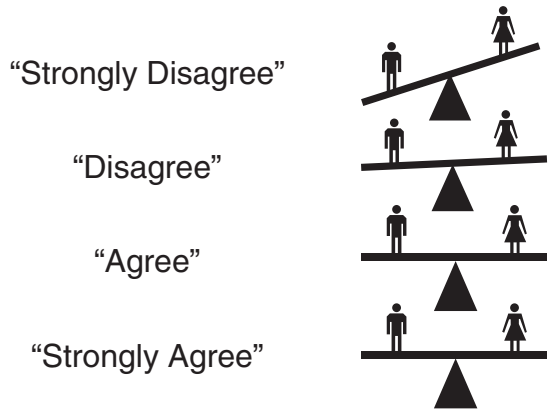



**Figure 6** An Illustration of Lower IDIF for Polytomous Outcomes

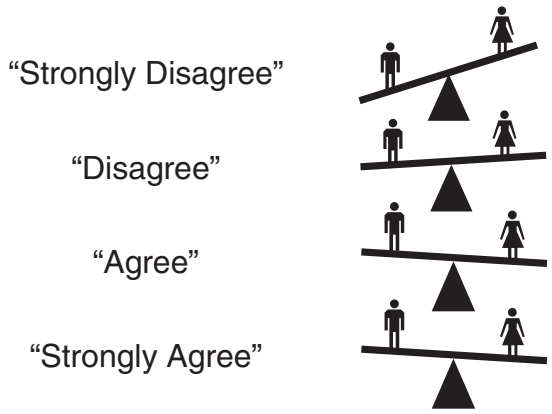



**Figure 7** An Illustration of Balanced IDIF for Polytomous Outcomes

## Assessment

The actual assessment and measurement of DIF is not always as straightforward as the concept. Various methods have been proposed to measure DIF. Perhaps the oldest method was an analysis of variance approach, which tested for an interaction effect between groups and items. Yet this approach did not gain in popularity because of the problematic nature of items being measured qualitatively or yielding binary outcomes. In 1972, William H. Angoff introduced one of the first widely used measures of DIF in the delta-plot method, also known as the transformed item-difficulty method. However, this method was often criticized as giving misleading results for items with differing discriminating power. Various other methods were introduced, such as the Mantel–Haenszel procedure. The Mantel–Haenszel procedure dominated the psychometric approach to the study of DIF for many years because of its ability to give an effect size for DIF, known as $\alpha$, in addition to a significance test.

Another approach to DIF analysis is based on IRT principles. While traditional methods allow for items to differ in difficulty, there is no allowance for differing item discrimination. As Angoff has stressed, it is possible for an item with the same difficulty parameter in the two groups but with different slope parameters to yield a DIF index of zero when analyzed by all but the IRT method. Thus, many IRT approaches to DIF emerged, in particular the multidimensional IRT method of Robin Shealy and William Stout in the early 1990s. The Shealy and Stout method provided an interesting vantage point for DIF analysis—that DIF could be a result of multidimensionality of the test in question. One criticism of the traditional methods is that they explain very little of the source of DIF. While the IRT perspective allows for a greater discernment of DIF than do traditional methods, there is still no attempt to explain the basis for DIF.

One way of approaching this issue in recent research is by using multilevel analysis techniques, such as the approach proposed by David B. Swanson, Brian E. Clauser, Susan M. Case, Ronald J. Nungster, and Carol Featherman. Kamata and Salih Binici, in 2003, considered a multilevel approach to DIF detection. In this model, Level 1 represented the item level, Level 2 represented the individual level, and Level 3 represented a group unit. The rationale for the inclusion of a third level was that the magnitude of DIF could vary across group units, such as schools in an educational DIF study. This approach models a random-effect DIF for the group units and uses individual characteristics to explicate the potential sources of DIF. Saengla Chaimongkol and Vaughn have, separately, extended the work of Kamata and Binici by using a Bayesian approach to obtain parameter estimates for dichotomously and polytomously scored responses, respectively.

*Brandon K. Vaughn*

*See also* Item Analysis; Item Response Theory

**Further Readings**

Angoff, W. H. (1972, September). *A technique for the investigation of cultural differences.* Paper presented at the annual meeting of the American Psychological Association, Honolulu, HI.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum.

Kamata, A., & Binici, S. (2003). *Random-effect DIF analysis via hierarchical generalized linear models.* Paper presented at the annual meeting of the Psychometric Society, Sardinia, Italy.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum.

Williams, V. S. L. (1997). The "unbiased" anchor: Bridging the gap between DIF and item bias. *Applied Measurement in Education, 10,* 253–267.

# Directional Hypothesis

A directional hypothesis is a prediction made by a researcher regarding a positive or negative change, relationship, or difference between two variables of a population. This prediction is typically based on past research, accepted theory, extensive experience, or literature on the topic. Key words that distinguish a directional hypothesis are: *higher, lower, more, less, increase, decrease, positive,* and *negative*. A researcher typically develops a directional hypothesis from research questions and uses statistical methods to check the validity of the hypothesis.

## Examples of Directional Hypotheses

A general format of a directional hypothesis would be the following: For (Population A), (Independent Variable 1) will be higher than (Independent Variable 2) in terms of (Dependent Variable). For example, "For ninth graders in Central High School, test scores of Group 1 will be higher than test scores of Group 2 in terms of Group 1 receiving a specified treatment." The following are other examples of directional hypotheses:

- There is a *positive* relationship between the number of books read by children and the children's scores on a reading test.
- Teenagers who attend tutoring sessions will make *higher* achievement test scores than comparable teenagers who do not attend tutoring sessions.

## Nondirectional and Null Hypotheses

In order to fully understand a directional hypothesis, there must also be a clear understanding of a nondirectional hypothesis and null hypothesis.

### *Nondirectional Hypothesis*

A nondirectional hypothesis differs from a directional hypothesis in that it predicts a change, relationship, or difference between two variables but does not specifically designate the change, relationship, or difference as being positive or negative. Another difference is the type of statistical test that is used. An example of a nondirectional hypothesis would be the following: For (Population A), there will be a difference between (Independent Variable 1) and (Independent Variable 2) in terms of (Dependent Variable 1). The following are other examples of nondirectional hypotheses:

- There is *a* relationship between the number of books read by children and the children's scores on a reading test.
- Teenagers who attend tutoring sessions will have achievement test scores that are *significantly different* from the scores of comparable teenagers who do not attend tutoring sessions.

### *Null Hypothesis*

Statistical tests are not designed to test a directional hypothesis or nondirectional hypothesis, but rather a null hypothesis. A null hypothesis is a prediction that there will be no change, relationship, or difference between two variables. A null hypothesis is designated by $H_o$. An example of a null hypothesis would be the following: for (Population A), (Independent Variable 1) will not be different from (Independent Variable 2) in terms of (Dependent Variable). The following are other examples of null hypotheses:

- There is *no* relationship between the number of books read by children and the children's scores on a reading test.
- Teenagers who attend tutoring sessions will make achievement test scores that are *equivalent* to those of comparable teenagers who do not attend tutoring sessions.

## Statistical Testing of Directional Hypothesis

A researcher starting with a directional hypothesis will have to develop a null hypothesis for the purpose for running statistical tests. The null hypothesis predicts that there will not be a change or relationship between variables of the two groups or populations. The null hypothesis is designated by $H_0$, and a null hypothesis statement could be written as $H_0 : \mu_1 = \mu^2$ (Population or Group 1 equals Population or Group 2 in terms of the dependent variable). A directional hypothesis or nondirectional hypothesis would then be considered to be an *alternative hypothesis* to the null hypothesis and would be designated as $H_1$. Since the directional hypothesis is predicting a direction of change or difference, it is designated as $H_1 : \mu_1 > \mu_2$ or $H_1 : \mu_1 < \mu_2$ (Population or Group 1 is greater than or less than Population or Group 2 in terms of the dependent variable). In the case of a nondirectional hypothesis, there would be no specified direction, and it could be designated as $H_1 : \mu_1 \neq \mu_2$ (Population or Group 1 does not equal Population or Group 2 in terms of the dependent variable).

When one is performing a statistical test for significance, the null hypothesis is tested to determine whether there is any significant amount of change, difference, or relationship between the two variables. Before the test is administered, the researcher chooses a significance level, known as an alpha level, designated by $\alpha$. In studies of education, the alpha level is often set at .05 or $\alpha = .05$. A statistical test of the appropriate variable will then produce a $p$ value, which can be understood as the probability a value as large as or larger than the statistical value produced by the statistical test would have been found by chance if the null hypothesis were true. The $p$ value must be smaller than the predetermined alpha level to be considered statistically significant. If no significance is found, then the null hypothesis is accepted. If there is a significant amount of change according to the $p$ value between two variables which cannot be explained by chance, then the null hypotheses is rejected, and the alternative hypothesis is accepted, whether it is a directional or a nondirectional hypothesis.

The type of alternative hypothesis, directional or nondirectional, makes a considerable difference in the type of significance test that is run. A nondirectional hypothesis is used when a two-tailed test of significance is run, and a directional hypothesis when a one-tailed test of significance is run. The reason for the different types of testing becomes apparent when examining a graph of a normalized curve, as shown in Figure 1.

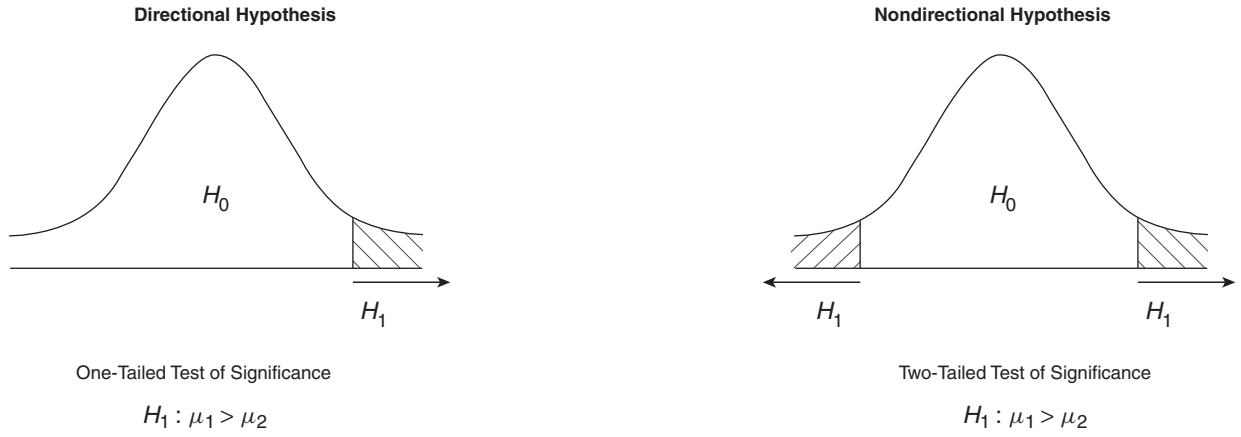



**Figure 1** Comparison of Directional and Nondirectional Hypothesis Test

The nondirectional hypothesis, since it predicts that the change can be greater or lesser than the null value, requires a two-tailed test of significance. On the other hand, the directional hypothesis in Figure 1 predicts that there will be a significant change greater than the null value; therefore, the negative area of significance of the curve is not considered. A one-tailed test of significance is then used to test a directional hypothesis.

### Summary Examples of Hypothesis Type

The following is a back-to-back example of the directional, nondirectional, and null hypothesis. In reading professional articles and test hypotheses, one can determine the type of hypothesis as an exercise to reinforce basic knowledge of research.

*Directional Hypothesis:* Women will have higher scores than men will on Hudson's self-esteem scale.

*Nondirectional Hypothesis:* There will be a difference by gender in Hudson's self-esteem scale scores.

*Null Hypothesis:* There will be no difference between men's scores and women's scores on Hudson's self-esteem scale.

*Ernest W. Brewer and Stephen Stockton*

***See also*** Alternative Hypotheses; Nondirectional Hypotheses; Null Hypothesis; One-Tailed Test; ***p*** Value; Research Question; Two-Tailed Test

### Further Readings

Ary, D., Jacobs, L. C., Razavieh, A., & Sorensen, C. (2006). *Introduction to research in education* (7th ed.). Belmont, CA: Thomson Wadsworth.

Borg, W. R. (1987). *Applying educational research* (2nd ed.). White Plains, NY: Longman.

Creswell, J. W. (2005). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research*. Upper Saddle River, NJ: Pearson Education.

Fraenkel, J. R., & Wallen, N. E. (2009). *How to design and evaluate research in education* (9th ed.). New York: McGraw-Hill.

Gay, L. R., Mills, G. E., & Airasian, P. (2009). *Educational research: Competencies for analysis and applications* (9th ed.). Upper Saddle River, NJ: Pearson Education.

Moore, D. S., & McMabe, G. P. (1993). *Introduction to the practice of statistics* (2nd ed.). New York: W. H. Freeman.

Patten, M. L. (1997). *Understanding research methods: An overview of the essentials*. Los Angeles: Pyrczak.

## Discourse Analysis

*Discourse* is a broadly used and abstract term that is used to refer to a range of topics in various disciplines. For the sake of this discussion, *discourse analysis* is used to describe a number of approaches to analyzing written and spoken language use beyond the technical pieces of language, such as words and sentences. Therefore, discourse analysis focuses on the use of language within a social context. Embedded in the constructivism–structuralism traditions, discourse analysis's key emphasis is on the use of language in social context. Language in this case refers to either text or talk, and context refers to the social situation or forum in which the text or talk occurs. Language and context are the two essential elements that help distinguish the two major approaches employed by discourse analysts. This entry discusses the background and major approaches of discourse analysis and frameworks associated with sociopolitical discourse analysis.

### Background

In the past several years social and applied or professional sciences in academia have seen a tremendous increase in the number of discourse analysis studies. The history of discourse analysis is long and embedded in the origins of a philosophical tradition of hermeneutics and phenomenology. These traditions emphasize the issue of *Verstehen,* or lifeworld, and the social interaction within the lifeworld. A few major theorists in this tradition are Martin Heidegger, Maurice Merleau-Ponty, Edmund Husserl, Wilhelm Dilthey, and Alfred Schutz. Early applications of discourse analysis in social and applied

and professional sciences can be found in psychology, sociology, cultural studies, and linguistics. The tradition of discourse analysis is often listed under interpretive qualitative methods and is categorized by Thomas A. Schwandt with hermeneutics and social construction under the constructivist paradigm. Jaber F. Gubrium and James A. Holstein place phenomenology in the same vein as naturalistic inquiry and ethnomethodology. The strong influence of the German and French philosophical traditions in psychology, sociology, and linguistics has made this a common method in the social and applied and professional sciences. Paradigmatically, discourse analysis assumes that there are multiple constructed realities and that the goal of researchers working within this perspective is to understand the interplay between language and social context. Discourse analysis is hermeneutic and phenomenological in nature, emphasizing the lifeworld and meaning making through the use of language. This method typically involves an analytical process of deconstructing and critiquing language use and the social context of language usage.

## Two Major Approaches

Discourse analysis can be divided into two major approaches: *language-in-use* (or socially situated text and talk) and *sociopolitical*. The language-in-use approach is concerned with the micro dimensions of language, grammatical structures, and how these features interplay within a social context. Language-in-use discourse analysis focuses on the rules and conventions of talk and text within a certain a context. This approach emphasizes various aspects of language within social context. Language-in-use methodologists focus on language and the interplay between language and social context. Language-in-use is often found in the disciplines of linguistics and literature studies and is rarely used in social and human sciences.

The second major approach, sociopolitical, is the focus of the rest of this entry because it is most commonly used within the social and human sciences. This approach is concerned with how language forms and influences the social context. Sociopolitical discourse analysis focuses on the social construction of discursive practices that maintain the social context. This approach emphasizes social context as influenced by language. Sociopolitical methodologists focus on social context and the interplay between social context and language. This approach is most often found in the social and professional and applied sciences, where researchers using sociopolitical discourse analysis often employ one of two specific frameworks: Foucaldian discourse analysis and critical discourse analysis (CDA).

## Sociopolitical Discourse Analysis Frameworks

### *Foucauldian Discourse Analysis*

Michel Foucault is often identified as the key figure in moving discourse analysis beyond linguistics and into the social sciences. The works of Foucault emphasize the sociopolitical approach to discourse analysis. Foucault emphasizes the role of discourse as power, which shifted the way discourse is critically analyzed. Foucault initially identified the concept of archeology as his methodology for analyzing discourse. Archeology is the investigation of unconsciously organized artifacts of ideas. It is a challenge to the present-day conception of history, which is a history of ideas. Archeology is not interested in establishing a timeline or Hegelian principles of history as progressive. One who applies archeology is interested in discourses, not as signs of a truth, but as the discursive practices that construct objects of knowledge. Archeology identifies *how* discourses of knowledge objects, separated from a historical-linear progressive structure, are formed. Therefore, archeology becomes the method of investigation, contradictory to the history of ideas, used when looking at an object of knowledge; archeology locates the artifacts that are associated with the discourses that form objects of knowledge. Archeology is the *how* of Foucauldian discourse analysis of the formation of an object of knowledge. Archeology consists of three key elements: *delimitation of authority* (who gets to speak about the object of knowledge?), *surface of emergence* (when does discourse about an object of knowledge begin?),

and *grids of specification* (how the object of knowledge is described, defined, and labeled).

However, Foucault's archeology then suggests a power struggle within the emergence of one or more discourses, via the identification of *authorities of delimitation*. Archeology's target is to deconstruct the history of ideas. The only way to fully deconstruct the history of an idea is to critique these issues of power. Hence, the creation of genealogy, which allows for this critique of power, with use of archeology, becomes the method of analysis for Foucault. Foucault had to create a concept like genealogy, since archeology's implied power dynamic and hints of a critique of power are in a form of hidden power. The term *genealogy* refers to the power relations rooted in the construction of a discourse. Genealogy focuses on the emergence of a discourse and identifies where power and politics surface in the discourse. Genealogy refers to the union of erudite knowledge and local memories, which allows us to establish a historical knowledge of struggles and to make use of this knowledge tactically today. Genealogy focuses on local, discontinuous, disqualified, illegitimate knowledge opposed to the assertions of the tyranny of totalizing discourses. Genealogy becomes the way we analyze the power that exists in the subjugated discourses that we find through the use of archeology. So genealogy is the exploration of the power that develops the discourse, which constructs an object of knowledge. The three key elements of genealogy include *subjugated discourses* (whose voices were minimized or hidden in the formation of the object of knowledge?), *local beliefs and understandings* (how is the object of knowledge perceived in the social context?), and *conflict and power relations* (where are the discursive disruptions and the enactments of power in the discourse?). Archeology suggests that there is a type of objectivity that indicates a positivistic concept of neutrality to be maintained when analyzing data. While genealogy has suggestions of subjectivity, localisms, and critique, much like postmodernist or critical theory, archeology focuses on *how* discourses form an object of knowledge. Genealogy becomes focused on *why* certain discourses are dominant in constructing an object of knowledge. Therefore, archeology is the method of data collection, and genealogy is the critical analysis of the data. These two concepts are not fully distinguishable, and a genealogy as Foucault defines it cannot exist without the method of archeology. Foucualt's work is the foundation of much of the sociopolitical discourse analysis used in contemporary social and applied and professional sciences. Many discourse studies cite Foucault as a methodological influence or use specific techniques or strategies employed by Foucault.

### *CDA*

CDA builds on the critique of power highlighted by Foucault and takes it a step further. Teun A.van Dijk has suggested that the central focus of CDA is the role of discourse in the (re)production and challenge of dominance. CDA's emphasis on the role of discourse in dominance specifically refers to social power enacted by elites and institutions' social and political inequality through discursive forms. The production and (re)production of discursive formation of power may come in various forms of discourse and power relations, both subtle and obvious. Therefore, critical discourse analysts focus on social structures and discursive strategies that play a role in the (re)production of power. CDA's critical perspective is influenced not only by the work of Foucault but also by the philosophical traditions of critical theorists, specifically Jurgen Habermas.

Norman Fairclough has stated that discourse is shaped and constrained by social structure and culture. Therefore he proposes three central tenets of CDA: social structure (class, social status, age, ethnic identity, and gender); culture (accepted norms and behaviors of a society); and discourse (the words and language we use). Discourse (the words and language we use) shapes our role and engagement with power within a social structure. CDA emphasizes when looking at discourse three levels of analysis: the text, the discursive practice, and the sociocultural practice.The text is a record of a communicated event that reproduces social power. Discursive practices are *ways of being in the world* that signify accepted social roles and identities. Finally, the sociocultural comprises the distinct context

where discourse occurs. The CDA approach attempts to link text and talk with the underlying power structures in society at a sociopolitical level through discursive practices. Text and talk are the description of communication that occurs within a social context that is loaded with power dynamics and structured rules and practices of power enactment. When text is not critically analyzed, oppressive discursive practices, such as marginalization and oppression, are taken as accepted norms. Therefore, CDA is intended to shine a light on such oppressive discursive practices. Discourse always involves power, and the role of power in a social context is connected to the past and the current context, and can be interpreted differently by different people due to various personal backgrounds, knowledge, and power positions. Therefore there is not one correct interpretation, but a range of appropriate and possible interpretations. The correct critique of power is not the vital point of CDA, but the process of critique and its ability to raise consciousness about power in social context is the foundation of CDA.

*Bart Miles*

**Further Readings**

Fairclough, N. (2000). *Language and power* (2nd ed.). New York: Longman.

Foucault, M. (1972). *The archaeology of knowledge* (A. M. Sheridan Smith, Trans.). London: Tavistock.

Schwandt, T. (2007). Judging interpretations. *New Directions for Evaluation, 114,* 11–15.

Stevenson, C. (2004). Theoretical and methodological approaches in discourse analysis. *Nurse Researcher, 12*(2), 17–29.

Taylor, S. (2001). Locating and conducting discourse analytic research. In M. Wetherell, S. Taylor, and S. J. Yates (Eds.), *Discourse as data: A guide for analysis* (pp. 5–48). London: Sage.

van Dijk, T. A. (1999). Critical discourse analysis and conversation analysis. *Discourse & Society, 10*(4), 459–460.

# Discriminant Analysis

Discriminant analysis is a multivariate statistical technique that can be used to predict group membership from a set of predictor variables. The goal of discriminant analysis is to find optimal combinations of predictor variables, called *discriminant functions,* to maximally separate previously defined groups and make the best possible predictions about group membership. Discriminant analysis has become a valuable tool in social sciences as discriminant functions provide a means to classify a case into the group that it mostly resembles and help investigators understand the nature of differences between groups. For example, a college admissions officer might be interested in predicting whether an applicant, if admitted, is more likely to succeed (graduate from the college) or fail (drop out or fail) based on a set of predictor variables such as high school grade point average, scores on the Scholastic Aptitude Test, age, and so forth. A sample of students whose college outcomes are known can be used to create a discriminant function by finding a linear combination of predictor variables that best separates Groups 1 (students who succeed) and 2 (students who fail). This discriminant function can be used to predict the college outcome of a new applicant whose actual group membership is unknown. In addition, discriminant functions can be used to study the nature of group differences by examining which predictor variables best predict group membership. For example, which variables are the most powerful predictors of group membership? Or what pattern of scores on the predictor variables best describes the differences between groups? This entry discusses the data considerations involved in discriminant analysis, the derivation and interpretation of discriminant functions, and the process of classifying a case into a group.

## Data Considerations of Discriminant Analysis

First of all, the predictor variables used to create discriminant functions must be measured at the interval or ratio level of measurement. The shape of the distribution of each predictor variable should correspond to a univariate normal distribution. That is, the frequency distribution of each predictor variable should be approximately bell shaped. In addition, *multivariate normality* of predictor variables is assumed in testing the significance of discriminant functions and calculating probabilities of group membership. The assumption of multivariate normality is met when each

variable has a univariate normal distribution at any fixed values of all other variables. Although the assumption of multivariate normality is complicated, discriminant analysis is found to be relatively robust with respect to the failure to meet the assumption if the violation is not caused by outliers. Discriminant analysis is very sensitive to the inclusion of outliers. Therefore, outliers must be removed or transformed before data are analyzed.

Another assumption of discriminant analysis is that no predictor variable may be expressed as a linear combination of other predictor variables. This requirement intuitively makes sense because when a predictor variable can be represented by other variables, the variable does not add any new information and can be considered redundant. Mathematically, such redundancy can lead to unreliable matrix inversions, which result in large standard errors of estimates. Therefore, redundant predictor variables must be excluded from the analysis.

A further assumption made in discriminant analysis is that the population variance–covariance matrices are equal across groups. This assumption is called *homogeneity of variance–covariance matrices*. When sample sizes are large or equal across groups, the significance test of discriminant function is usually robust with respect to the violation of the homogeneity assumption. However, the classification is not so robust in that cases tend to be overclassified into groups with greater variability. When sample sizes are small and unequal, the failure to meet the homogeneity assumption often causes misleading results of both significance tests and classifications. Therefore, prior to performing discriminant analysis, the tenability of the assumption of homogeneity of variance–covariance matrices must be tested.

## Deriving and Testing Discriminant Functions

The major task in discriminant analysis is to find discriminant functions that maximally separate groups and make the best possible predictions about group membership. The easiest and most commonly used type of discriminant function is the linear function, in which predictor variables are weighted and then summed to produce the discriminant score. The equation of this type of discriminant function has the following form:

$$D_i = d_{i0} - d_{i1}X_1 - d_{i2}X_2 - \cdots - d_{iK}X_K,$$

where $X_1, X_2, \ldots, X_K$ are predictor variables 1 to $K, D_i$ is the $i$th discriminant function, $d_{i0}$ is a constant, and $d_{iK}$ is the coefficient of predictor variable $k$ for discriminant function $i$. Discriminant functions are like regression equations in the sense that a discriminant score for each case is predicted by multiplying the score on each predictor variable ($X_k$) by its associated coefficient ($d_{iK}$), summing over all predictors, and adding a constant $d_{i0}$. When there are only two groups in discriminant analysis, only one discriminant function is needed to best separate groups. However, when there are more than two groups, the maximum number of discriminant functions that can be derived is equal to the number of groups minus one or the number of predictor variables, whichever is fewer. For example, for a discriminant analysis with three groups and four predictor variables, two discriminant functions can be derived. The weights or coefficients of predictor variables for the first function can be derived so that the group means on the function are as different as possible. The weights used to combine predictor variables in the second function can be determined also based on the criterion of producing maximum possible difference among group means but with the additional condition that the values of the second function are not correlated with values of the first function. Similarly, the third function also maximally differentiates among groups but with the restriction of being uncorrelated with the first two functions, and so forth.

Although multiple discriminant functions are often identified, not all functions significantly discriminate among groups. Therefore, statistical tests are needed to determine the significance of each discriminant function. There are several statistics for testing the significance of each discriminant function, such as Wilks's lambda, Roy's largest root, Hotelling's trace, and Pillai's criterion. Wilks's lambda is the ratio of within-groups variance to total variance and therefore represents the percentage of variance in discriminant scores not explained by group membership. Wilks's lambda ranges from 0 to 1, with values closer to 1 indicating the function is less discriminating. Because Wilks's lambda can be

transformed to a statistic that has a chi-square distribution, its statistical significance can be tested. A significant Wilks's lambda indicates that the group means calculated from the discriminant analysis are significantly different and therefore the discriminant function works well in discriminating among groups.

## Interpreting Discriminant Functions

If a discriminant function is found to be significant, one might be interested in discovering how groups are separated along the discriminant function and which predictor variables are most useful in separating groups. To visually inspect how well groups are spaced out along discriminant function, individual discriminant scores and group centroids can be plotted along the axes formed by discriminant functions. The mean of discriminant scores within a group is known as the group centroid. If the centroids of two groups are well separated and there is no obvious overlap of the individual cases along a discriminant function, then the discriminant function separates the two groups well. If group centroids are close to each other and individual cases overlap a great deal, the discriminant function fails to provide a clear separation of the two groups. When there are only one or two significant discriminant functions, the location of group centroids and data cases can be easily plotted. However, when there are more than two discriminant functions, it will be visually difficult to locate the group centroids and data cases. Therefore, only pairwise plots of discriminant functions are used. However, the plot based on the first two discriminant functions is expected to be most informative because these two functions are the most powerful discriminators of groups.

The relative importance of the contribution of each variable to the separation of the groups can be evaluated by examining the discriminant coefficients. At this stage, however, the discriminant function can be considered an unstandardized equation in the sense that raw scores on predictor variables are used to produce discriminant scores. Although the magnitudes of unstandardized coefficients indicate the absolute contribution of a predictor variable in determining the discriminant score, this information might be misleading when one attempts to evaluate the relative importance of predictor variables. This is because when standard deviations are not the same across predictor variables, one unit change in the value of a variable varies from one variable to another. Therefore, standardized discriminant coefficients are needed. Standardized discriminant coefficients indicate the relative importance of measured variables in calculating discriminant scores. Standardized discriminant coefficients involve adjusting the unstandardized discriminant coefficients by the variance of the raw scores on each predictor variable. Standardized discriminant coefficients would be obtained if the original data were converted to standard form, in which each variable has a mean of zero and standard deviation of one, and then used to optimize the discriminant coefficients. However, the standardized discriminant coefficients can be derived from unstandardized coefficients directly by the following formula:

$$s_{ik} = d_{ik}\sqrt{\frac{\mathrm{SS}_k}{N-g}},$$

where $s_{ik}$ and $d_{ik}$ are the standardized and unstandardized coefficients for predictor variable $k$ on discriminant function $i$, respectively, $SS_k$ is the sum of squares associated with variable $k$, $N$ is the total number of cases, and $g$ is the number of groups. The predictor variable associated with the largest standardized coefficient (in absolute value) contributes most to determining scores on the discriminant function and therefore plays the most important role in separating groups. It should be noted that when predictor variables are correlated, the associated discriminating coefficients might provide misleading results. For example, consider two correlated variables that have rather small contributions to the discriminant function. The two estimated standardized coefficients might be large but with opposite signs, so that the effect of one variable is, to some degree, canceled by the opposite effect of the other variable. However, this could be misinterpreted as both variables having relatively large contributions to the discriminant function but in different directions.

A better guide to the meaning of the discriminant function is to use structure coefficients. Structure coefficients look at the correlations between the discriminant function and each predictor variable. The variable that correlates most highly with the discriminant function shares the greatest amount of variance with the discriminant function and, therefore, explains the discriminant function more. Structure coefficients can be directly derived by calculating the correlation coefficients between each of the predictor variables and the discriminant scores. It addresses the question, To which of the $K$ variables is the discriminant function most closely related? When the absolute value of the coefficient is very large (close to 1), the discriminant function is carrying nearly the same information as the predictor variable. In comparison, when the coefficient is near zero, the discriminant function and the predictor variable share little variance. The discriminant function can be named after the predictor variables that have the highest correlations.

## Classifications

In discriminant analysis, discriminant functions can be used to make predictions of the group to which a case most likely belongs. Classification of an individual case involves calculation of the individual's discriminant score and comparison of it with each of the group centroids. To make predictions of group membership, the distance from the individual's discriminant scores to each of the group centroids is measured, and the centroid to which the individual's scores are closest is the group to which the individual is predicted to belong. A distance measure commonly used in discriminant analysis is Mahalanobis $D^2$, which calculates the squared distance from a specific case to each of the group centroids. $D^2$ can be considered a measure that represents the degree to which a case's profile on the predictor variables resembles the typical profile of a group. Based on this interpretation, a case should be classified into the group with the smallest $D^2$. Because $D^2$ is a statistic with a chi-square distribution of $p$ degrees of freedom, where $p$ is the number of predictor variables, the probabilities that a case belongs to a group can be calculated. Similar to $D^2$, these probabilities can also be used to assign group membership. That is, a case will be classified into the group to which it has the highest probability of belonging. In addition, the probabilities of group membership serve as an indicator of the discriminating power of discriminant functions. For example, discriminant functions are considered to function well when a case has a high probability of belonging to one group but low probabilities of belonging to other groups. In this way, it is clear that the case should be classified into the group of the highest probability. However, if probabilities for all groups are very close, it might be meaningless to classify the case into a specific group given that groups are actually not very distinct based on the discriminant functions.

When predicted group membership is compared with actual group membership in the sample from which the function was calculated, the percentage of correct predictions, often called the *hit ratio,* can be calculated. To evaluate the performance of classification, the hit ratio should not be compared with zero but rather with the percentage that would have been correctly classified by chance. If the groups have equal sample sizes, the expected percentage of correct predictions by chance is equal to $1/K$, where $K$ is the total number of groups. For instance, for a two-group analysis with equal sample sizes, one can expect a 50% of chance of making correct predictions of group membership by pure random guesses, and therefore the expected hit ratio based on chance is .5. If the hit ratio yielded by discriminant functions is .6, the improvement is actually rather small. When groups are unequal in size, the percentage that could be correctly classified by chance can be estimated by multiplying the expected probabilities of each group by the corresponding group size, summing for all groups, and dividing the sum by the total sample size. A $z$ test for the difference between proportions can be performed to statistically test the significance of the improvement in the classification accuracy from the discriminant analysis.

It should be noted that the hit ratio tends to overestimate the classification accuracy of discriminant functions when the same sample is used to both derive the discriminant function and test its predictive ability. To overcome this,

the effectiveness of the classification procedure can be tested on another sample that is independent of the one used to derive discriminant function. Testing the classification procedure on a new sample is called cross-validation. There are several methods of cross-validation, including the use of a *holdout* sample, *double cross-validation,* and the so-called *jackknife* procedure. In the holdout method of cross-validation, the sample is randomly split into two subsets. One is used to develop discriminant functions and the other is used to test the accuracy of the classification procedure. This is an unbiased method of estimating the true misclassification rate. However, large sample sizes are required. The idea of double cross-validation is similar to the use of a holdout sample for cross-validation. In double cross-validation, the total sample is divided in half. Separate discriminant analyses are performed on each sample, and the results are cross-validated on the other sample. In the jackknife procedure, one observation at a time is eliminated from the sample, the discriminant analysis is performed with the remaining observations, and then the obtained discriminant functions are used to classify the eliminated observation. This process is repeated until all observations have been eliminated once. Consequently, the classification rates can be determined using the cumulative results.

*Ying Cui*

***See also*** Canonical Correlation Analysis; Logistic Regression; Multiple Regression; Multivariate Analysis of Variance (MANOVA)

**Further Readings**

Huberty, C. J. (1984). Issues in the use and interpretation of discriminant analysis. *Psychological Bulletin, 95,* 156–171.

Klecka, W. R. (1980). *Discriminant analysis*. Beverly Hills, CA: Sage.

Spicer, J. (2005). *Making sense of multivariate data analysis*. Thousand Oaks, CA: Sage.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston: Pearson Education.

Timm, N. H. (2002). *Applied multivariate analysis*. New York: Springer.

# Discussion Section

The purpose of a discussion section of a research paper is to relate the results (results section) back to the initial hypotheses of the study (introduction section). The discussion section provides an interpretation of the results, presents conclusions, and supports all the conclusions with evidence from the study and generally accepted knowledge. The discussion section should describe (a) what new knowledge has been gained in the study and (b) where research should go next.

Discussion of research findings must be based on basic research concepts. When the researcher explains a phenomenon, he or she must explain mechanisms. If the researcher's results agree with the expectations, the researcher should describe the theory that the evidence supported. If the researcher's results differ from the expectations, then he or she should explain why that might have happened. If the researcher cannot make a decision with confidence, he or she should explain why that was the case, and how the study might be modified in the future. Because one study will not answer an overall question, the researcher, keeping the big picture in mind, addresses where research goes next.

Ideally, the scientific method discovers cause-and-effect relationships between variables (conceptual objects whose value may vary). An independent (exposure) variable is one that, when changed, causes a change in another variable, the dependent (outcome) variable. However, a change in a dependent variable may be due wholly or in part to a change in a third, confounding (extraneous) variable. A confounding variable is anything other than the independent variable of interest that may affect the dependent variable. It must be predictive of the outcome variable independent of its association with the exposure variable of interest, but it cannot be an intermediate in the causal chain of association between exposure and outcome. Confounding variables can be dealt with through the choice of study design and/or data analysis.

Bias is the systematic deviation of results or inferences from the truth, or processes leading to such deviation. Validity refers to the lack of bias, or the credibility of study results and the degree to

which the results can be applied to the general population of interest. Internal validity refers to the degree to which conclusions drawn from a study correctly describe what actually transpired during the study. External validity refers to whether and to what extent the results of a study can be generalized to a larger population (the target population of the study from which the sample was drawn, and other populations across time and space).

Threats to validity include *selection bias* (which occurs in the design stage of a study), *information bias* (which occurs in the data collection stage of a study), and *confounding bias* (which occurs in the data analysis stage of a study). Selection bias occurs when during the selection step of the study, the participants in the groups to be compared are not comparable because they differ in extraneous variables other than the independent variable under study. In this case, it would be difficult for the researcher to determine whether the discrepancy in the groups is due to the independent variable or to the other variables. Selection bias affects internal validity. Selection bias also occurs when the characteristics of subjects selected for a study are systematically different from those of the target population. This bias affects external validity. Selection bias may be reduced when group assignment is randomized (in experiments) or selection processes are controlled for (in observational studies). Information bias occurs when the estimated effect is distorted either by an error in measurement or by misclassifying the participant for independent (exposure) and/or dependent (outcome) variables. In experiments, information bias may be reduced by improving the accuracy of measuring instruments and by training technicians. In observational studies, information bias may be reduced by pretesting questionnaires and training interviewers. Confounding bias occurs when statistical controlling techniques (stratification or mathematical modeling) are not used to adjust for the effects of confounding variables. Therefore, a distorted estimate of the exposure effect results because the exposure effect is mixed with the effects of extraneous variables. Confounding bias may be reduced by performing a "dual" analysis (with and without adjusting for extraneous variables). Although adjusting for confounders ensures unbiasedness, unnecessary adjustment for nonconfounding variables always reduces the statistical power of a study. Therefore, if both results in a dual analysis are similar, then the unadjusted result is unbiased and should be reported based on power considerations. If both results are different, then the adjusted one should be reported based on validity considerations.

Below is a checklist of the items to be included in a discussion section:

1. *Overview:* Provide a brief summary of the most important parts of the introduction section and then the results section.
2. *Interpretation:* Relate the results back to the initial study hypotheses. Do they support or fail to support the study hypotheses? It is also important to discuss how the results relate to the literature cited in the introduction. Comment on the importance and relevance of the findings and how the findings are related to the big picture.
3. *Strengths and limitations:* Discuss the strengths and limitations of the study.
4. *Recommendations:* Provide recommendations on the practical use of current study findings and suggestions for future research.

The following are some tips for researchers to follow in writing the discussion section: (a) Results do not prove hypotheses right or wrong. They support them or fail to provide support for them. (b) In the case of a correlation study, causal language should not be used to discuss the results. (c) Space is valuable in scientific journals, so being concise is imperative. Some journals ask authors to restrict discussion to four pages or less, double spaced, typed. That works out to approximately one printed page. (d) When referring to information, data generated by the researcher's own study should be distinguished from published information. Verb tense is an important tool for doing that—past tense can be used to refer to work done; present tense can be used to refer to generally accepted facts and principles.

The discussion section is important because it interprets the key results of a researcher's study in light of the research hypotheses under study and the published literature. It should provide a good indication of what the new findings from the

researcher's study are and where research should go next.

*Bernard Choi and Anita Pak*

***See also*** Bias; Methods Section; Results Section; Validity of Research Conclusions

**Further Readings**

Branson, R. D. (2004). Anatomy of a research paper. *Respiratory Care, 49,* 1222–1228.

Choi, P. T. (2005). Statistics for the reader: What to ask before believing the results. *Canadian Journal of Anesthesia, 52,* R1-R5.

Hulley, S. B., Newman, T. B., & Cummings, S. R. (1988). The anatomy and physiology of research. In S. B. Hulley & S. R. Cummings (Eds.), *Designing clinical research* (pp. 1–11). Baltimore: Williams & Wilkins.

# Dissertation

As a requirement for an advanced university degree, the dissertation is usually the last requirement a candidate fulfills for a doctorate. Probably its most salient characteristic is that it is a unique product, one that embodies in some way the creativity of the author—the result of research and of original thinking and the creation of a physical product. Depending on departmental tradition, some dissertations are expected to be solely originated by the candidate; in others, the topic (and sometimes the approach as well) is given by the major professor. But even in the latter case, the candidates are expected to add something of their own originality to the end result.

This description of some relatively common features of the dissertation requirement applies primarily to higher education in the United States. That there are common features owes much to communication among universities, no doubt through such agencies as the Council of Graduate Schools and the American Association of Universities. But the requirement's evolution at the local level has resulted in considerable variation across the differing cultures of universities and even the departments within them.

As a means of maintaining high standards, many universities administer their doctoral program through a graduate school with its own dean. Additionally, some universities designate certain faculty, those who have proven they are researchers, as graduate faculty who participate in setting advanced degree policies and serve as major professors and chairs. This dual faculty status has disappeared at most universities, however.

The dissertation process typically moves through three stages: the proposal stage; the activation stage, in which the research, thinking, or producing work is accomplished; and the final stage of presentation and approval. Though distinguishable for explanatory purposes, these stages are often blurred in practice. This is particularly evident when the area in which one intends to work is known, but not the specific aspect. For example, the proposal and activation stage often merge until the project outlines become clear.

In most cases one faculty member from the department serves as the major professor or committee chair (henceforth referred to as the *chair*). This is usually at the invitation of the student, although some departments assign chairs in order to equitably balance faculty load. Additional faculty are recruited by the student to serve as readers or committee members, often at the suggestion of the chair. Dissertation chairpersons and committee members are chosen for their experience in the candidate's topic of interest and/or for some special qualifications, such as experience with the research method or knowledge of statistics or experimental design.

## The Proposal Stage

Depending on the department's tradition, the dissertation may or may not be a collaborative affair with the faculty. Regardless, the dissertation, beginning with the formulation of the problem in the proposal, is often a one-on-one, give-and-take relation between the candidate and the committee chair. In *How to Prepare a Dissertation Proposal,* David R. Krathwohl and Nick L. Smith described a dissertation proposal as a logical plan of work to learn something of real or potential significance about an area of interest. Its opening problem statement draws the reader into the plan: showing its significance, describing how it builds on previous work

(both substantively and/or methodologically), and outlining the investigation. The whole plan of action flows from the problem statement: the activities described in the design section, their sequence often illuminated graphically in the work plan (and, if one is included, by the time schedule), and their feasibility shown by the availability of resources. Krathwohl and Smith point out that a well-written proposal's enthusiasm should carry the reader along and reassure the reader with its technical and scholarly competence. A solid proposal provides the reader with such a model of the clarity of thought and writing to be expected in the final write-up that the reader feels this is an opportunity to support research that should not be missed.

While at first it may appear that this definition suggests that the proposal should be written like an advertisement, that is not what it is intended to convey. It simply recognizes the fact that if students cannot be enthusiastic about their idea, it is a lot to expect others to be. Material can be written in an interesting way and still present the idea with integrity. It doesn't have to be boring to be good.

Second, the definition points out that the proposal is an integrated chain of reasoning that makes strong logical connections between the problem statement and the coherent plan of action the student has proposed undertaking.

Third, this process means that students use this opportunity to present their ideas and proposed actions for consideration in a shared decision making situation. With all the integrity at their command, they help their chair or doctoral committee see how they view the situation, how the idea fills a need, how it builds on what has been done before, how it will proceed, how pitfalls will be avoided, why pitfalls not avoided are not a serious threat, what the consequences are likely to be, and what significance they are likely to have.

Fourth, while the students' ideas and action plans are subject to consideration, so also is their capability to successfully carry them through.

Such a proposal definition gives the student a goal, but proposals serve many purposes besides providing an argument for conducting the study and evidence of the student's ability. Proposals also serve as a request for faculty commitment, as a contract, as an evaluative criterion, and as a partial dissertation draft.

Faculty members who assume the role of dissertation chair take on a substantial commitment of time, energy, and in some instances resources. Agreeing to be the student's chair usually involves a commitment to help where able, such as in procuring laboratories, equipment, participants, access to research sites, and funding. Thus, nearly all faculty set limits on how many doctoral candidates they will carry at any one time.

In cases in which students take on problems that are outside the interests of any departmental faculty, students may experience difficulty in finding a faculty member to work with them because of the substantial additional time commitment and the burden of gaining competence in another area. If no one accepts them, the students may choose to change topic or, in some instances, transfer to another university.

Few view the proposal as a binding contract that if fulfilled, will automatically lead to a degree. Nevertheless, there is a sense that the proposal serves as a good faith agreement whereby if the full committee approves the proposal and the student does what is proposed with sufficient quality (whatever that standard means in the local context), then the student has fulfilled his or her part of the bargain, and the faculty members will fulfill theirs. Clearly, as an adjunct to its serving as a contract, the proposal also serves as an evaluative criterion for "fulfilling his or her part."

In those institutions in which there is a formal admission to candidacy status, the proposal tends to carry with it more of a faculty commitment: The faculty have deemed the student of sufficient merit to make the student a candidate; therefore the faculty must do what they can to help the student successfully complete the degree.

Finally, the proposal often becomes part of the dissertation itself. The format for many dissertations is typically five chapters:

1. Statement of the problem, why it is of some importance, and what one hopes to be able to show
2. A review of the past research and thinking on the problem, how it relates to what the student intends to do, and how this project builds on it and possibly goes beyond it—substantively and methodologically

3. The plan of action (what, why, when, how, where, and who)
4. What was found, the data and its processing
5. Interpretation of the data in relation to the problem proposed.

Many departments require students to prepare the proposal as the first three chapters to be used in the dissertation with appropriate modification.

It is often easier to prepare a proposal when the work to be done can be preplanned. Many proposals, however, especially in the humanities and more qualitatively oriented parts of the social sciences, are for emergent studies. The focus of work emerges as the student works with a given phenomenon. Without a specific plan of work, the student describes the study's purpose, the approach, the boundaries of the persons and situations as well as rules for inclusion or exclusion, and expected findings. Since reality may be different, rules for how much deviation requires further approval are appropriate.

Practice varies, but all institutions require proposal approval by the chair, if not the whole committee. Some institutions require very formal approval, even an oral examination on the proposal; others are much looser.

## The Activation Phase

This phase also varies widely in how actively the chair and committee monitor or work with the student. Particularly where faculty members are responsible for a funded project supporting a dissertation, monitoring is an expected function. But for many, just how far faculty members are expected or desired to be involved in what is supposed to be the student's own work is a fine line. Too much and it becomes the professor's study rather than the student's. Most faculties let the students set the pace and are available when called on for help.

Completion appears to be a problem in all fields; it is commonly called the all-but-dissertation (ABD) problem. Successful completion of 13 to 14 years of education and selection into a doctoral program designates these students as exceptional, so for candidates to fail the final hurdle is a waste of talent. Estimates vary and differ by subject matter, but sciences have the highest completion rates after 10 years of candidacy—70% to 80%—and English and the humanities the lowest—30%. The likelihood of ABD increases when a student leaves campus before finishing. Reasons vary, but finances are a major factor. Michael T. Nettles and Catherine M. Millett's Rate of Progress scale allows students to compare their progress with that of peers in the same field of study. The PhD Completion Project of the Council of Graduate Schools aims to find ways of decreasing ABD levels.

## The Final Stage

Many dissertations have a natural stopping point: the proposed experiment concludes, one is no longer learning anything new, reasonably available sources of new data are exhausted. The study is closed with data analysis and interpretation in relation to what one proposed. But if one is building a theoretical model, developing or critiquing a point of view, describing a situation, or developing a physical product, when has one done enough? Presumably when the model is adequately described, the point of view appropriately presented, or the product works on some level. But "adequate," "appropriate," and "some level" describe judgments that must be made by one's chair and committee, and their judgment may differ from that of the student.

The time to face the decision of how much is enough is as soon as the research problem is sufficiently well described that criteria that the chair and committee deem reasonable can be ascribed to it—a specified period of observations or number of persons to be queried, certain books to be digested and brought to bear, and so forth. While not a guaranteed fix because minds change as the problem emerges, the salience of closure conditions is always greater once the issue is raised.

Dissertations are expected to conform to the standard of writing and the appropriate style guide for their discipline. In the social sciences this guide is usually either American Psychological Association or Modern Language Association style.

Once the chair is satisfied (often also the committee), the final step is, in most cases, an oral examination. Such examinations vary in their

formality, who chairs (often chosen by the graduate school), and how large the examining body. The examiners usually include the committee, other faculty from the department, and at least one faculty member from other departments (called outsiders), schools, or colleges and sometimes other universities (also often chosen by the graduate school). Other students may silently observe.

The chair (usually an outsider but in some universities the committee chair) usually begins with an executive session to set the procedure (often traditional but sometimes varied). Then the candidate and any visitors are brought back in and the candidate presents, explains, and, if necessary, defends his or her choices. The chair rotates the questioning, usually beginning with the outsiders and ending with committee members. When questioning is finished (often after about 2 hours), the examiners resume executive session to decide dissertation acceptance: as is, with minor modifications, with major modifications, or rejection. Rejection is rare; minor modifications are the norm.

### Accessibility of Dissertations

Most universities require that students provide a copy of their dissertation in suitable format for public availability. A copy (print or digital, depending on the library's preference) is given to the university's library. Copies are usually also required to be filed with ProQuest UMI Dissertation Publishing and/or the Electronic Thesis/Dissertation Open Access Initiative (OAI) Union Catalog of the Networked Digital Library of Theses and Dissertation (NDLTD).

*David R. Krathwohl*

***See also*** American Psychological Association Style; Proposal

#### Further Readings

Krathwohl, D. R., & Smith, N. L. (2005). *How to prepare a dissertation proposal.* Syracuse, NY: Syracuse University Press.

Nettles, M. T., & Millett, C. M. (2006). *Three magic letters: Getting to Ph.D.* Baltimore: Johns Hopkins University Press.

#### Websites

Electronic Thesis/Dissertation Open Access Initiative (OAI) Union Catalog of the Networked Digital Library of Theses and Dissertation (NDLTD): http://www.ndltd.org/find

ProQuest UMI Dissertation Publishing: http://www.proquest.com/products_rumi/dissertations

## Distribution

A distribution refers to the way in which researchers organize sets of scores for interpretation. The term also refers to the underlying probabilities associated with each possible score in a real or theoretical population. Generally, researchers plot sets of scores using a curve that allows for a visual representation. Displaying data in this way allows researchers to study trends among scores. There are three common approaches to graphing distributions: histograms, frequency polygons, and ogives. Researchers begin with a set of raw data and ask questions that allow them to characterize the data by a specific distribution. There are also a variety of statistical distributions, each with its unique set of properties.

Distributions are most often characterized by whether the data are discrete or continuous in nature. Dichotomous or discrete distributions are most commonly used with nominal and ordinal data. A commonly discussed discrete distribution is known as a binomial distribution. Most textbooks use the example of a coin toss when discussing the probability of an event occurring with two discrete outcomes. With continuous data, researchers most often examine the degree to which the scores approximate a normal distribution. Because distributions of continuous variables can be normal or nonnormal in nature, researchers commonly explain continuous distributions in one of four basic ways: *average value, variability, skewness,* and *kurtosis*. The normal distribution is often the reference point for examining a distribution of continuously scored variables. Many continuous variables have distributions that are bell shaped and are said to approximate a normal distribution. The theoretical curve, called the *bell curve,* can be used to study many variables that are not normally

distributed but are approximately normal. According to the *central limit theorem,* as the sample size increases, the shape of the distribution of the sample means taken will approach a normal distribution.

## Discrete Distributions

The most commonly discussed discrete probability distribution is the *binominal distribution.* The binomial distribution is concerned with scores that are dichotomous in nature, that is, there can be only one of two possible outcomes. The *Bernoulli trial* (named after mathematician Jakob Bernoulli) is a good example that is often used when teaching students about a binomial distribution of scores. The most often discussed Bernoulli trial is that of flipping a coin, in which the outcome will be either heads or tails. The process allows for estimating the probability that an event will occur. Binomial distributions can also be used when one wants to determine the probability associated with correct or incorrect responses. In this case, an example might be a 10-item test that is scored dichotomously (correct/incorrect). A binomial distribution allows us to calculate the probability of scoring 5 out of 10, 6 out of 10, 7 out of 10, and so on, correct. Because the calculation of binomial probability distributions can become somewhat tedious, binomial distribution tables often accompany many statistics textbooks so that researchers can quickly access information regarding such estimates. It should be noted that binomial distributions are most often used in nonparametric procedures. Chi-square distributions are another form of a discrete distribution that is often used when one wants to report whether an expected outcome occurred due to chance alone.

## Continuous Distributions

Continuous variables can be any value or interval associated with a number line. In theory, a continuous variable can assume an infinite number of possible values with no gaps among the intervals. This is sometimes referred to as a "smooth" process. To graph a continuous probability distribution, one draws a horizontal axis that represents the values associated with the continuous variable. Above the horizontal axis is drawn a curve that encompasses all values within the distribution. The area between the curve and the horizontal axis is sometimes referred to as the *area under the curve.* Generally speaking, distributions that are based on continuous data tend to cluster around an average score, or *measure of central tendency.* The measure of central tendency that is most often reported in research is known as the *mean.* Other measures of central tendency include the *median* and the *mode.* Bimodal distributions occur when there are two modes or two values that occur most often in the distribution of scores. The median is often used when there are extreme values at either end of the distribution. When the mean, median, and mode are the same value, the curve tends to be bell shaped, or normal, in nature. This is one of the unique features of what is known as the normal distribution. In such cases, the curve is said to be *symmetrical about the mean,* which means that the shape is the same on both sides. In other words, if one drew a perpendicular line through the mean score, each side of the curve would be a perfect reflection the other. *Skewness* is the term used to measure a lack of symmetry in a distribution. Skewness occurs when one tail of the distribution is longer than the other. Distributions can be positively or negatively skewed depending on which tail is longer. In addition, distributions can differ in the amount of variability. *Variability* explains the dispersion of scores around the mean. Distributions with considerable dispersion around the mean tend to be flat when compared to the normal curve. Distributions that are tightly dispersed around the mean tend to be peaked in nature when compared to the normal curve with the majority of scores falling very close to the mean. In cases in which the distribution of scores appears to be flat, the curve is said to be *platykurtic,* and distributions that are peaked compared with the normal curve are said to be *leptokurtic* in nature. The flatness or peakedness of a distribution is a measure of kurtosis, which, along with variability and skewness, helps explain the shape of a distribution of scores. Each normally distributed variable will have its own measure of central tendency, variability, degree of skewness, and kurtosis. Given this fact, the shape and location of the curves will vary for many normally distributed variables. To avoid needing to have a table of areas under the curve for each

normally distributed variable, statisticians have simplified things through the use of the standard normal distribution based on a *z*-score metric with a mean of zero and a standard deviation of 1. By standardizing scores, we can estimate the probability that a score will fall within a certain region under the normal curve. Parametric test statistics are typically applied to data that approximate a normal distribution, and *t* distributions and *F* distributions are often used. As with the binomial distribution and chi-square distribution, tables for the *t* distribution and *F* distribution are typically found in most introductory statistics textbooks. These distributions are used to examine the variance associated with two or more sets of sample means. Because the sampling distribution of scores may vary based on the sample size, the calculation of both the *t* and *F* distributions includes something called *degrees of freedom,* which is an estimate of the sample size for the groups under examination. When a distribution is said to be nonnormal, the use of nonparametric or distribution-free statistics is recommended.

*Vicki Schmitt*

***See also*** Bernoulli Distribution; Central Limit Theorem; Frequency Distribution; Kurtosis; Nonparametric Statistics; Normal Distribution; Parametric Statistics

**Further Readings**

Bluman, A. G. (2009). *Elementary statistics: A step by step approach* (7th ed.). Boston: McGraw-Hill.

Howell, D. C. (2010). *Statistical methods for psychology* (7th ed.). Belmont, CA: Wadsworth Cengage Learning.

Larose, D. T. (2010). *Discovering statistics*. New York: Freeman.

Salkind, N. J. (2008). *Statistics for people who (think they) hate statistics* (3rd ed.). Thousand Oaks, CA: Sage.

# Disturbance Terms

In the field of research design, researchers often want to know whether there is a relationship between an observed variable (say, $y$) and another observed variable (say, $x$). To answer the question, researchers may construct the model in which $y$ depends on $x$. Although $y$ is not necessarily explained only by $x$, a discrepancy always exists between the observed value of $y$ and the predicted value of $y$ obtained from the model. The discrepancy is taken as a *disturbance term* or an *error term*.

Suppose that $n$ sets of data, $(x_1, y_1)$, $(x_2, y_2)$, $\ldots$, $(x_n, y_n)$, are observed, where $y_i$ is a scalar and $x_i$ is a vector (say, $1 \times k$ vector). We assume that there is a relationship between $x$ and $y$, which is represented as the model $y = f(x)$, where $f(x)$ is a function of $x$. We say that $y$ is explained by $x$, or $y$ is *regressed on* $x$. Thus $y$ is called the dependent or explained variable, and $x$ is a vector of the independent or explanatory variables. Suppose that a vector of the unknown parameter (say $\beta$, which is a $k \times 1$ vector) is included in $f(x)$. Using the $n$ sets of data, we consider estimating $\beta$ in $f(x)$. If we add a *disturbance term* (say $u$, which is also called an *error term*), we can express the relationship between $y$ and $x$ as $y = f(x) + u$. The disturbance term $u$ indicates the term that cannot be explained by $x$. Usually, $x$ is assumed to be nonstochastic. Note that $x$ is said to be nonstochastic when it takes a fixed value. Thus $f(x)$ is *deterministic,* while $u$ is *stochastic*. The researcher must specify $f(x)$. Representatively, it is often specified as the linear function $f(x) = x\beta$.

The reasons a disturbance term $u$ is necessary are as follows: (a) There are some unpredictable elements of randomness in human responses, (b) an effect of a large number of omitted variables is contained in $x$, (c) there is a measurement error in $y$, or (d) a functional form of $f(x)$ is not known in general. Corresponding examples are as follows: (a) Gross domestic product data are observed as a result of human behavior, which is usually unpredictable and is thought of as a source of randomness. (b) We cannot know all the explanatory variables that depend on $y$. Most of the variables are omitted, and only the important variables needed for analysis are included in $x$. The influence of the omitted variables is thought of as a source of $u$. (c) Some kinds of errors are included in almost all the data, either because of data collection difficulties or because the explained variable is inherently unmeasurable, and a proxy variable has to be used in their stead. (d) Conventionally we

specify $f(x)$ as $f(x) = x\beta$. However, there is no reason to specify the linear function. Exceptionally, we have the case in which the functional form of $f(x)$ comes from the underlying theoretical aspect. Even in this case, however, $f(x)$ is derived from a very limited theoretical aspect, not every theoretical aspect.

For simplicity hereafter, consider the linear regression model $y_i = x_i\beta + u_i, \quad i = 1,2,\ldots,n$. When $u_1, u_2, \ldots, u_n$ are assumed to be mutually independent and identically distributed with mean zero and variance $\sigma^2$, the sum of squared residuals,

$$\sum_{i=1}^{n} (y_i - x_i\beta)^2,$$

is minimized with respect to $\beta$. Then, the estimator of $\beta$ (say, $\hat{\beta}$) is

$$\hat{\beta} = \left(\sum_{i=1}^{n} x_i'x_i\right)^{-1} \sum_{i=1}^{n} x_i'y_i,$$

which is called the *ordinary least squares* (OLS) *estimator*. $\hat{\beta}$ is known as the best linear unbiased estimator (BLUE). It is distributed as

$$N\left(\beta, \sigma^2\left(\sum_{i=1}^{n} x_i'x_i)^{-1}\right)\right)$$

under the normality assumption on $u_i$, because $\hat{\beta}$ is rewritten as

$$\hat{\beta} = \beta + \left(\sum_{i=1}^{n} x_i'x_i\right)^{-1} \sum_{i=1}^{n} x_i'u_i.$$

Note from the central limit theorem that $\sqrt{n}(\hat{\beta} - \beta)$ is asymptotically normally distributed with mean zero and variance $\sigma^2 M_{xx}^{-1}$ even when the disturbance term $u_i$ is not normal, in which case we have to assume $(1/n)\sum_{i=1}^{n} x_i'x_i \to M_{xx}$ as $n$ goes to infinity; that is, $n \to \infty$ ($a \to b$ indicates that $a$ approaches $b$).

For the disturbance term $u_i$, we have made the following three assumptions:

1. $V(u_i) = \sigma^2$ for all $i$,
2. $Cov(u_i, u_j) = 0$ for all $i \neq j$,
3. $Cov(u_i, x_j) = 0$ for all $i$ and $j$.

Now we examine $\hat{\beta}$ in the case in which each assumption is violated.

## Violation of the Assumption $V(u_i) = \sigma^2$ For All $i$

When the assumption on variance of $u_i$ is changed to $V(u_i) = \sigma_i^2$, that is, a heteroscedastic disturbance term, the OLS estimator $\hat{\beta}$ is no longer BLUE. The variance of $\hat{\beta}$ is given by

$$V(\hat{\beta}) = (\sum_{i=1}^{n} x_i'x_i)^{-1}(\sum_{i=1}^{n} \sigma_i^2 x_i'x_i) (\sum_{i=1}^{n} x_i'x_i)^{-1}.$$

Let $b$ be a solution of minimization of $\sum_{i=1}^{n} (y_i - x_i\beta)^2/\sigma_i^2$ with respect to $\beta$. Then

$$b = (\sum_{i=1}^{n} x_i'x_i/\sigma_i^2)^{-1} \sum_{i=1}^{n} x_i'y_i/\sigma_i^2$$

and

$$b \sim N(\beta, (\sum_{i=1}^{n} x_i'x_i/\sigma_i^2)^{-1})$$

are derived under the normality assumption for $u_i$. We have the result that $\hat{\beta}$ is not BLUE because of $V(b) \leq V(\hat{\beta})$. The equality holds only when $\sigma_i^2 = \sigma^2$ for all $i$. For estimation, $\sigma_i^2$ has to be specified, such as $\sigma_i = |z_i\gamma|$, where $z_i$ represents a vector of the other exogenous variables.

## Violation of the Assumption $Cov(u_i, u_j) = 0$ For All $i \neq j$

The correlation between $u_i$ and $u_j$ is called the *spatial correlation* in the case of cross-sectional data and the *autocorrelation* or *serial correlation* in time-series data. Let $\rho_{ij}$ be the correlation coefficient between $u_i$ and $u_j$, where $\rho_{ij} = 1$ for all $i = j$ and $\rho_{ij} = \rho_{ji}$ for all $i \neq j$. That is, we have $Cov(u_i, u_j) = \sigma^2\rho_{ij}$. The matrix that the $(i,j)$th element is $\rho_{ij}$ should be positive definite. In this situation, the variance of $\hat{\beta}$ is:

$$V(\hat{\beta}) = \sigma^2(\sum_{i=1}^{n} x_i'x_i)^{-1}(\sum_{i=1}^{n}\sum_{j=1}^{n} \rho_{ij}x_i'x_j) \times (\sum_{i=1}^{n} x_i'x_i)^{-1}.$$

Let $b^*$ be a solution of the minimization problem of $\sum_{i=1}^{n}\sum_{j=1}^{n} \rho^{ij}(y_i - x_i\beta)'(y_j - x_j\beta)$ with respect to $\beta$, where $\rho^{ij}$ denotes the $(i,j)$th element of the inverse matrix of the matrix that the $(i,j)$th

element is $\rho_{ij}$. Then, under the normality assumption on $u_i$, we obtain

$$b^* = (\sum_{i=1}^n \sum_{j=1}^n \rho^{ij} x_i' x_j)^{-1} (\sum_{i=1}^n \sum_{j=1}^n \rho^{ij} x_i' y_j)$$

and

$$b^* \sim N(\beta, \sigma^2 (\sum_{i=1}^n \sum_{j=1}^n \rho^{ij} x_i' x_j)^{-1}).$$

It can be verified that we obtain the following: $V(b^*) \leq V(\hat{\beta})$. The equality holds only when $\rho_{ij} = 1$ for all $i = j$ and $\rho_{ij} = 0$ for all $i \neq j$. For estimation, we need to specify $\rho_{ij}$. For an example, we may take the following specification: $\rho_{ij} = \rho^{|i-j|}$, which corresponds to the first-order autocorrelation case (i.e., $u_i = \rho u_{i-1} + \varepsilon_i$, where $\varepsilon_i$ is the independently distributed error term) in time-series data. For another example, in the spatial correlation model we may take the form $\rho_{ij} = 1$ when $i$ is in the neighborhood of $j$ and $\rho_{ij} = 0$ otherwise.

### Violation of the Assumption $Cov(u_i, x_j) = 0$ For All $i$ and $j$

If $u_i$ is correlated with $x_j$ for some $i$ and $j$, it is known that $\hat{\beta}$ is not an unbiased estimator of $\beta$, that is, $E(\hat{\beta}) \neq \beta$, because of $E((\sum_{i=1}^n x_i' x_i)^{-1} x \sum_{i=1}^n x_i' u_i) \neq 0$. In order to obtain a consistent estimator of $\beta$, we need the condition $(1/n) \sum_{i=1}^n x_i' u_i \to 0$ as $n \to \infty$. However, we have the fact that $(1/n) \sum_{i=1}^n x_i' u_i \not\to 0$ as $n \to \infty$ in the case of $Cov(u_i, x_j) \neq 0$. Therefore, $\hat{\beta}$ is not a consistent estimator of $\beta$, that is, $\hat{\beta} \not\to \beta$ as $n \to \infty$. To improve this inconsistency problem, we use the instrumental variable (say, $z_i$), which satisfies the properties $(1/n) \sum_{i=1}^n z_i' u_i \to 0$ and $(1/n) \sum_{i=1}^n z_i' x_i \not\to 0$ as $n \to \infty$. Then, it is known that $b_{IV} = (\sum_{i=1}^n z_i' x_i)^{-1} \sum_{i=1}^n z_i' y_i$ is a consistent estimator of $\beta$, that is, $b_{IV} \to \beta$ as $n \to \infty$. Therefore, $b_{IV}$ is called the *instrumental variable* estimator. It can be also shown that $\sqrt{n}(b_{IV} - \beta)$ is asymptotically normally distributed with mean zero and variance $\sigma^2 M_{zx}^{-1} M_{zz} M_{xz}^{-1}$, where

$$(1/n) \sum_{i=1}^n z_i' x_i \to M_{zx},$$

$$(1/n) \sum_{i=1}^n z_i' z_i \to M_{zz},$$

and

$$(1/n) \sum_{i=1}^n x_i' z_i \to M_{xz} = M_{zx}'$$

as $n \to \infty$. As an example of $z_i$, we may choose $z_i = \hat{x}_i$, where $\hat{x}_i$ indicates the predicted value of $x_i$ when $x_i$ is regressed on the other exogenous variables associated with $x_i$, using OLS.

*Hisashi Tanizaki*

***See also*** Autocorrelation; Central Limit Theorem; Serial Correlation; Unbiased Estimator

**Further Readings**

Kennedy, P. (2008). *A guide to econometrics* (6th ed.). Malden, MA: Wiley-Blackwell.

Maddala, G. S. (2001). *Introduction to econometrics* (3rd ed.). New York: Wiley.

## *Doctrine of Chances, The*

*The Doctrine of Chances*, by Abraham de Moivre, is frequently considered the first textbook on probability theory. Its subject matter is suggested by the book's subtitle, namely, *A Method of Calculating the Probabilities of Events in Play.* Here "play" signifies games involving dice, playing cards, lottery draws, and so forth, and the "events" are specific outcomes, such as throwing exactly one ace in four throws.

De Moivre was a French Protestant who escaped religious persecution by emigrating to London. There he associated with some of the leading English scientists of the day, including Edmund Halley and Isaac Newton (to whom *The Doctrine of Chances* was dedicated). At age 30 de Moivre was elected to the Royal Society and much later was similarly honored by scientific academies in both France and Prussia. Yet because he never succeeded in procuring an academic position, he was obliged to earn a precarious livelihood as a private tutor, teacher, and consultant and died in severe poverty. These adverse circumstances seriously constrained the amount of time he could devote to original research. Even so, de Moivre not only made substantial contributions to probability

theory but also helped found analytical trigonometry, discovering a famous theorem that bears his name.

To put *The Doctrine of Chances* in context, and before discussing its contributions and aftermath, it is first necessary to provide some historical background.

## Antecedents

De Moivre was not unusual in concentrating the bulk of his book on games of chance. This emphasis was apparent from the very first work on probability theory. The mathematician Gerolamo Cardano, who was also a professional gambler, wrote *Liber de ludo aleae* (Book on Games of Chance), in which he discussed the computation of probabilities. However, because Cardano's work was not published until 1663, the beginning of probability theory is traditionally assigned to 1654. In that year Blaise Pascal and Pierre de Fermat began a correspondence on gaming problems. This letter exchange led Pascal to write *Traité du triangle arithmétique* (Treatise on the Arithmetical Triangle), in which he arranged the binomial coefficients into a triangle and then used them to solve certain problems in games of chance. De Moivre was evidently among the first to refer to this geometric configuration as *Pascal's triangle* (even though Pascal did not really introduce the schema).

In 1657 Christian Huygens published *Libellus de ratiociniis in ludo aleae* (The Value of All Chances in Games of Fortune), an extended discussion of certain issues raised by Pascal and Fermat. Within a half century, Huygens's work was largely superseded by two works that appeared shortly before de Moivre's book. The first was the 1708 *Essai d'analyse sur les jeux de hasard* (Essay of Analysis on Games of Chance) by Pierre de Montmort and *Ars Conjectandi* (The Art of Conjecturing) by Jakob (or James) Bernoulli, published posthumously in 1713. By the time de Moivre wrote *The Doctrine of Chances,* he was familiar with all these efforts as well as derived works. This body of knowledge put him in a unique position to create a truly comprehensive treatment of probability theory.

## Contributions

To be more precise, *The Doctrine of Chances* was actually published in four versions dispersed over most of de Moivre's adult life. The first version was a 52-page memoir that he had published in Latin in the *Philosophical Transactions of the Royal Society.* In 1711 this contribution was entitled "De Mensura Sortis" (On the Measurement of Lots). The primary influence was Huygens, but not long afterwards the author encountered the work of Montmort and Bernoulli. Montmort's work left an especially strong imprint when de Moivre expanded the article into the first edition of *The Doctrine of Chances,* which was published in 1718. In fact, the influence was so great that Montmort and de Moivre entered into a priority dispute that, fortunately, was amicably resolved (unlike the 1710 dispute between Newton and Leibniz over the invention of the calculus that de Moivre helped arbitrate in Newton's favor).

Sometime after the first edition, de Moivre began to pursue Bernoulli's work on approximating the terms of the binomial expansion. By 1733 de Moivre had derived what in modern terms is called the *normal approximation to the binomial distribution.* Although originally published in a brief Latin note, the derivation was translated into English for inclusion in the second edition of *The Doctrine of Chances* in 1738. This section was then expanded for the book's third edition, which was published posthumously in 1756. This last edition also includes de Moivre's separate contributions to the theory of annuities, plus a lengthy appendix. Hence, because the last version is by far the most inclusive in coverage, it is this publication on which his reputation mostly rests. In a little more than 50 years, the 52-page journal article had grown into a 348-page book—or 259 pages if the annuity and appendix sections are excluded.

The main part of the text is devoted to treating 74 problems in probability theory. Besides providing the solutions, de Moivre offers various cases and examples and specifies diverse corollaries and lemmas. Along the way the author presents the general procedures for the addition and multiplication of probabilities, discusses probability-generating functions, the binomial distribution law, and the use of recurring series

to solve difference equations involving probabilities, and offers original and less restricted treatments of the duration of play in games of chance (i.e., the "gambler's ruin" problem). Although some of the mathematical terminology and notation is archaic, with minor adjustments and deletions *The Doctrine of Chances* could still be used today as a textbook in probability theory. Because the author intended to add to his meager income by the book's sale, it was written in a somewhat more accessible style than a pure mathematical monograph.

Yet from the standpoint of later developments, the most critical contribution can be found on pages 243–254 of the 3rd edition (or pages 235–243 of the 2nd edition), which are tucked between the penultimate and final problems. It is here that de Moivre presented "a method of approximating the sum of the terms of the binomial $(a + b)^n$ expanded into a series, from whence are deduced some practical rules to estimate the degree of assent which is to be given to experiments" (put in modern mathematical notation and expressed in contemporary English orthography). Going beyond Bernoulli's work (and that of Nicholas Bernoulli, Jakob's nephew), the approximation is nothing other than the normal (or Gaussian) curve.

Although de Moivre did not think of the resulting exponential function in terms of a probability density function, as it is now conceived, he clearly viewed it as describing a symmetrical bell-shaped curve with inflection points on both sides. Furthermore, even if he did not possess the explicit concept of the standard deviation, which constitutes one of two parameters in the modern formula (the other being the mean), de Moivre did have an implicit idea of a distinct and fixed unit that meaningfully divided the curve on either side of the maximum point. By hand calculation he showed that the probabilities of outcomes coming within $\pm 1$, 2, and 3 of these units would be .6827, .9543, and .9987 (rounding his figures to four decimal places). The corresponding modern values for $\pm 1$, 2, and 3 standard deviations from the mean are .6826, .9544, and .9974. Taken together, de Moivre's understanding was sufficient to convince Karl Pearson and others to credit him with the original discovery of the normal curve.

There are two additional aspects of this work that are worth mentioning, even if not as important as the normal curve itself.

First, de Moivre established a special case of the central limit theorem that is sometimes referred to as the *theorem of de Moivre–Laplace*. In effect, the theorem states that as the number of independent (Bernoulli) trials increases indefinitely, the binomial distribution approaches the normal distribution. De Moivre illustrated this point by showing that a close approximation to the normal curve could be obtained simply by flipping a coin a sufficient number of times. This demonstration is basically equivalent to that of the bean machine or quincunx that Francis Galton invented to make the same point.

Second, de Moivre offered the initial components of what later became known as the Poisson approximation to the binomial distribution, albeit it was left to Siméon Poisson to provide this derivation the treatment it deserved. Given this fragmentary achievement and others, one can only imagine what de Moivre would have achieved had he obtained a chair of mathematics at a major European university.

## Aftermath

Like his predecessors Huygens, Montmort, and Jakob Bernoulli, de Moivre was primarily interested in what was once termed *direct* probability. That is, given a particular probability distribution, the goal was to infer the probability of a specified event. To offer a specific example, the aim was to answer questions such as What is the probability of throwing a score of 12 given three throws of a regular six-faced die? In contrast, these early mathematicians were not yet intrigued by problems in *inverse* probability. In this case the goal is to infer the underlying probability distribution that would most likely produce a set of observed events. An instance would be questions like, Given that 10 coin tosses yielded 6 heads and 4 tails, what is the probability that it is still an unbiased coin? and how many coin tosses would we need before we knew that the coin was unbiased with a given degree of confidence? Inverse probability is what we now call statistical inference—the inference of population properties from small random samples taken from that population. How can we infer the population

distribution from the sample distribution? How much confidence can we place in using the sample mean as the estimate of the population mean?

This orientation toward direct rather than inverse probability makes good sense historically. As already noted, probability theory was first inspired by games of chance. And such games begin with established probability distributions. That is how each game is defined. So a coin toss should have two equally likely outcomes, a die throw six equally likely outcomes, and a single draw from a full deck of cards, 52 equally likely outcomes. The probabilities of various compound outcomes—like getting one and only one ace in three throws—can therefore be derived in a direct and methodical manner. In these derivations one certainty (the outcome probability) is derived from another certainty (the prior probability distribution) by completely certain means (the laws of probability). By comparison, because inverse probability deals with uncertainties, conjectures, and estimates, it seems far more resistant to scientific analysis. It eventually required the introduction of such concepts as confidence intervals and probability levels.

It is telling that when Jakob Bernoulli attempted to solve a problem of the latter kind, he dramatically failed. He specifically dealt with an urn model with a given number of black and white pebbles. He then asked how many draws (with replacement) a person would have to make before the relative frequencies could be stated with an a priori level of confidence. After much mathematical maneuverings—essentially constituting the first power analysis—Bernoulli came up with a ludicrous answer: 25,550 observations or tests. Not surprisingly, he just ended *Ars Conjectandi* right there, apparently without a general conclusion, and left the manuscript unpublished at his death. Although de Moivre made some attempt to continue from where his predecessor left off, he was hardly more successful, except for the derivation of the normal curve.

It is accordingly ironic that the normal curve eventually provided a crucial contribution to statistical inference and analysis. First, Carl Friedrich Gauss interpreted the curve as a density function that could be applied to measurement problems in astronomy. By assuming that errors of measurement were normally distributed, Gauss could derive the method of least squares to minimize those errors. Then Pierre-Simon Laplace, while working on the central limit theorem, discovered that the distribution of sample means tends to be described by a normal distribution, a result that is independent of the population distribution. Adolphe Quételet later showed that human individual differences in physical characteristics could be described by the same curve. The average person (*l'homme moyen*) was someone who resided right in the middle of the distribution. Later still Galton extended this application to individual differences in psychological attributes and defined the level of ability according to placement on this curve. In due course the concept of univariate normality was generalized to those of bivariate and multivariate normality. The normal distribution thus became the single most important probability distribution in the behavioral and social sciences—with implications that went well beyond what de Moivre had more modestly envisioned in *The Doctrine of Chances*.

*Dean Keith Simonton*

***See also*** Game Theory; Probability, Laws of; Significance Level, Concept of; Significance Level, Interpretation and Construction

**Further Readings**

de Moivre, A. (1967). *The doctrine of chances* (3rd ed.). New York: Chelsea. (Original work published 1756)

Hald, A. (2003). *A history of probability and statistics and their applications before 1750*. Hoboken, NJ: Wiley.

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Cambridge, MA: Harvard University Press.

## Double-Blind Procedure

A double-blind procedure refers to a procedure in which experimenters and participants are "blind to" (without knowledge of) crucial aspects of a study, including the hypotheses, expectations, or, most important, the assignment of participants to experimental groups. This entry discusses the implementation and application of double-blind procedures, along with their historical background

and some of the common criticisms directed at them.

## Experimental Control

"Double-blinding" is intimately coupled to *randomization,* where participants in an experimental study are allocated to groups according to a random algorithm. Participants and experimenters are then blinded to group allocation. Hence double-blinding is an additional control element in experimental studies. If only some aspect of a study is blinded, it is a *single-blind* study. This is the case when the measurement of an outcome parameter is done by someone who does not know which group a participant belongs to and what hypotheses and expectations are being tested. This could, in principle, also be done in nonexperimental studies if, for instance, two naturally occurring cohorts, smokers and nonsmokers, say, are tested for some objective marker, such as intelligence or plasma level of hormones. Double-blinding presupposes that participants are allocated to the experimental procedure and control procedure at random. Hence, by definition, natural groups or cohorts cannot be subject to double-blinding. Double-blind testing is a standard for all pharmaceutical substances, such as drugs, but should be implemented whenever possible in all designs. In order for a study to succeed with double-blinding, a control intervention uses a placebo that can be manufactured in a way that makes the placebo indistinguishable from the treatment.

## Allocation Concealment and Blind Analysis

There are two corollaries to double-blinding: allocation concealment and blind statistical analysis. If an allocation algorithm, that is, the process of allocating participants to experimental groups, is completely random, then, by definition, the allocation of participants to groups is concealed. If someone were to allocate participants to groups in an alternating fashion, then the allocation would not be concealed. The reason is that if someone were to be unblinded, because of an adverse event, say, then whoever knew about the allocation system could trace back and forth from this participant and find out about the group allocation of the other participants.

Double-blind studies are normally also evaluated "blind." Here, the data are input by automatic means (Internet, scanning), or by assistants blind to group allocation of participants. Whatever procedures are done to prepare the database for analysis, such as transformations, imputations of missing values, and deletion of outliers, is done without knowledge of group assignment. Normally a study protocol stipulates the final statistical analysis in advance. This analysis is then run with a database that is still blinded in the sense that the groups are named "A" and "B." Only after this first and definitive analysis has been conducted and documented is the blind broken.

Good clinical trials also test whether the blinding was compromised during the trial. If, for instance, a substance or intervention has many and characteristic side effects, then patients or clinicians can often guess whether someone was allocated to treatment (often also called *verum,* from the Latin word for *true*) or placebo. To test for the integrity of the blinding procedure, either all participants or a random sample of them are asked, before the blind is broken, what group they think they had been allocated to. In a good, uncompromised blinded trial, there will be a near-random answer pattern because some patients will have improved under treatment and some under control.

## Placebos

To make blinding of patients and clinicians possible, the control procedure has to be a good mock or placebo procedure (also sometimes called *sham*). In pharmaceutical trials this is normally done by administering the placebo in a capsule or pill of the same color but containing pharmacologically inert material, such as corn flour. If it is necessary to simulate a taste, then often other substances or coloring that are inactive or only slightly active, such as vitamin C, are added. For instance, if someone wants to create a placebo for caffeine, quinine can be used. Sometimes, if a pharmacological substance has strong side effects, an active placebo might be used. This is a substance that produces some of the side effects, but hardly any of the desired pharmacological effects, as the experimental treatment.

If a new pharmacological substance is to be tested against an already existing one, then a *double-dummy* technique is used. Placebos for both substances are manufactured, and a patient takes always two substances, one of which is a placebo for the other substance.

## Blinding in Behavioral Research

While pharmaceutical procedures are comparatively easy to blind, behavioral, surgical, or other procedures are difficult to blind. For surgical and similar trials, often real blinds are used, that is, screens that shield part of the surgical team from seeing what is actually happening such that only a few operating surgeons know whether they are actually performing the real operation or a sham procedure. Thus, the rest of the team—anesthesiologist, nurses, assistant surgeons who might take performance measures after the operation—can remain blinded. However, it should be noted that most surgical procedures have not been evaluated by blinded trials.

Most behavioral interventions cannot be tested in blinded trials. A psychotherapist, for instance, has to know what he or she is doing in order to be effective. However, patients can be allocated in a blind fashion to different types of interventions, one of which is a sham intervention. Therefore, in behavioral research, active controls are very important. These are controls that are tailor made and contain some, but not all, the purportedly active elements of the treatment under scrutiny. The less possible it is to blind patients and clinicians to the interventions used, the more important it is to use either blinded assessors to measure the outcome parameter of interest or objective measures that are comparatively robust against experimenter or patient expectation.

## Historical Background

Historically, blinding was introduced in the testing of so-called animal magnetism introduced by Franz-Anton Mesmer, a German healer, doctor, and hypnotist. Mesmer thought that a magnet would influence a magnetic field in the human body that regulates human physiology. This purported specific human magnetic system he called *animal magnetism*. After emigrating to France, Mesmer set up an enormously successful practice, treating rich and poor alike with his magnetic cures. Several medical professors converted to animal magnetism, among them Charles D'Eslon, a professor in the medical school. The Académie Française decided to scrutinize the phenomenon in 1784. D'Eslon and a volunteer patient were tested. When the magnetiseur stroked the volunteer with his magnet, she had all sorts of fits and exhibited strong physical reactions. It was decided to place a curtain between her and the magnetiseur. This step showed that only if the participant could see her magnetiseur did the phenomenon occur with some reliability. By the same token, this proved that the participant's conscious or unconscious expectation, and not the magnet, conveyed the influence.

This was the birth of hypnosis as a scientific discipline as we know it. Incidentally, it was also the birth of blinded control in experiments. The next step in the history of blinded experimentation was taken by followers of Samuel Hahnemann, the German doctor who had invented homeopathy. Homeopaths use sugar globules impregnated with alcohol but containing hardly any or none of the pharmacological substances that are nominally their source, yet claim effects. This pharmacological intervention lends itself perfectly to blinded investigations because the placebo is completely undistinguishable from the experimental medicine. Homeopaths were the first to test their intervention in blinded studies, in 1834 in Paris, where inert sugar globules were dispensed with the normal homeopathic ritual, with roughly equal effectiveness. This was done again in 1842 in Nuremberg, where homeopaths challenged critics and dispensed homeopathic and inert globules under blind conditions to volunteers who were to record their symptoms. Symptoms of a quite similar nature were reported by both groups, leaving the battle undecided. The first blinded conventional medical trials were conducted in 1865 by Austin Flint in the United States and William Withey Gull in London, both testing medications for rheumatic fever. The year 1883 saw the first blinded psychological experiment. Charles Sanders Peirce and Joseph Jastrow wanted to know the smallest weight difference that participants could sense

and used a screen to blind the participants from seeing the actual weights. Blinded tests became standard in hypnosis and parapsychological research. Gradually medicine also came to understand the importance and power of suggestion and expectation. The next three important dates are the publication of *Methodenlehre der therapeutischen Untersuchung* (Clinical Research Methodology) by German pharmacologist Paul Martini in 1932, the introduction of randomization by Ronald Fisher's *Design of Experiments* in 1935, and the 1945 Cornell conferences on therapy, which codified the blinded clinical trial.

## Caveats and Criticisms

Currently there is a strong debate over how to balance the merits of strict experimental control with other important ingredients of therapeutic procedures. The double-blind procedure has grown out of a strictly mechanistic, pharmacological model of efficacy, in which only a single specific physiological mechanism is important, such as the blocking of a target receptor, or one single psychological process that can be decoupled from contexts. Such careful focus can be achieved only in strictly experimental research with animals and partially also with humans. But as soon as we reach a higher level of complexity and come closer to real-world experiences, such blinding procedures are not necessarily useful or possible. The real-world effectiveness of a particular therapeutic intervention is likely to consist of a specific, mechanistically active ingredient that sits on top of a variety of other effects, such as strong, nonspecific effects of relief from being in a stable therapeutic relationship; hope that a competent practitioner is structuring the treatment; and reduction of anxiety through the security given by the professionalism of the context. This approach has been discussed under the catchword *whole systems research,* which acknowledges (a) that a system or package of care is more than just the sum of all its elements and (b) that it is unrealistic to assume that all complex systems of therapy can be disentangled into their individual elements. Both pragmatic and theoretical reasons stand against it. Pragmatically speaking, blinded clinical trials are cost intensive, and researchers will likely not be able to muster the resources to run sufficient numbers of isolated, blinded trials on all components to gain enough certainty. Theoretically, therapeutic packages come in a bundle that falls apart if one were to disentangle them into separate elements. So care has to be taken not to overgeneralize the pharmacological model to all situations.

Blinding is always a good idea, where it can be implemented, because it increases the internal validity of a study. Double-blinding is necessary if one wants to know the specific effect of a mechanistic intervention.

*Harald Walach*

***See also*** Experimenter Expectancy Effect; Hawthorne Effect; Internal Validity; Placebo; Randomization Tests

### Further Readings

Committee for Proprietary Medicinal Products Working Party on Efficacy of Medicinal Products. (1995). Biostatistical methodology in clinical trials in applications for marketing authorizations for medicinal products. CPMP working party on efficacy of medicinal products note for guidance III/3630/92-EN. *Statistics in Medicine, 14,* 1659–1682.

Crabtree, A. (1993). *From Mesmer to Freud: Magnetic sleep and the roots of psychological healing.* New Haven, CT: Yale University Press.

Greenberg, R. P., Bornstein, R. F., Greenberg, M. D., & Fisher, S. (1992). A meta-analysis of antidepressant outcome under "blinder" conditions. *Journal of Consulting & Clinical Psychology, 60,* 664–669.

Kaptchuk, T. J. (1998). Intentional ignorance: A history of blind assessment and placebo controls in medicine. *Bulletin of the History of Medicine, 72,* 389–433.

Schulz, K. F., Chalmers, I., Hayes, R. J., & Altman, D. G. (1995). Empirical evidence of bias: Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *Journal of the American Medical Association, 273,* 408–412.

Shelley, J. H., & Baur, M. P. (1999). Paul Martini: The first clinical pharmacologist? *Lancet, 353,* 1870–1873.

White, K., Kando, J., Park, T., Waternaux, C., & Brown, W. A. (1992). Side effects and the 'blindability' of clinical drug trials. *American Journal of Psychiatry, 149,* 1730–1731.

# Dummy Coding

Dummy coding is used when categorical variables (e.g., sex, geographic location, ethnicity) are of interest in prediction. It provides one way of using categorical predictor variables in various kinds of estimation models, such as linear regression. Dummy coding uses only 1s and 0s to convey all the necessary information on group membership. With this kind of coding, the researcher enters a 1 to indicate that a person is a member of a category, and a 0 otherwise.

Dummy codes are a series of numbers assigned to indicate group membership in any mutually exclusive and exhaustive category. Category membership is indicated in one or more columns of 0s and 1s. For example, a researcher could code sex as 1 = *female,* 0 = *male* or 1 = *male,* 0 = *female.* In this case the researcher would have a column variable indicating status as male or female. In general, with $k$ groups there will be $k - 1$ coded variables. Each of the dummy-coded variables uses 1 degree of freedom, so $k$ groups have $k - 1$ degrees of freedom, just as in analysis of variance (ANOVA). Consider the following example, in which there are four observations within each of the four groups:

| *Group* | *G1* | *G2* | *G3* | *G4* |
|---|---|---|---|---|
| | 1 | 2 | 5 | 10 |
| | 3 | 3 | 6 | 10 |
| | 2 | 4 | 4 | 9 |
| | 2 | 3 | 5 | 11 |
| *Mean* | *2* | *3* | *5* | *10* |

For this example we need to create three dummy-coded variables. We will call them d1, d2, and d3. For d1, every observation in Group 1 will be coded as 1 and observations in all other groups will be coded as 0. We will code d2 with 1 if the observation is in Group 2 and zero otherwise. For d3, observations in Group 3 will be coded 1 and zero for the other groups. There is no d4; it is not needed because d1 through d3 have all the information needed to determine which observation is in which group.

Here is how the data look after dummy coding:

| *Values* | *Group* | *d1* | *d2* | *d3* |
|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 0 |
| 3 | 1 | 1 | 0 | 0 |
| 2 | 1 | 1 | 0 | 0 |
| 2 | 1 | 1 | 0 | 0 |
| 2 | 2 | 0 | 1 | 0 |
| 3 | 2 | 0 | 1 | 0 |
| 4 | 2 | 0 | 1 | 0 |
| 3 | 2 | 0 | 1 | 0 |
| 5 | 3 | 0 | 0 | 1 |
| 6 | 3 | 0 | 0 | 1 |
| 4 | 3 | 0 | 0 | 1 |
| 5 | 3 | 0 | 0 | 1 |
| 10 | 4 | 0 | 0 | 0 |
| 10 | 4 | 0 | 0 | 0 |
| 9 | 4 | 0 | 0 | 0 |
| 11 | 4 | 0 | 0 | 0 |

Note that every observation in Group 1 has the dummy-coded value of 1 for d1 and 0 for the others. Those in Group 2 have 1 for d2 and 0 otherwise, and for Group 3, d3 equals 1 with 0 for the others. Observations in Group 4 have all 0s on d1, d2, and d3. These three dummy variables contain all the information needed to determine which observations are included in which group. If you are in Group 2, then d2 is equal to 1 while d1 and d3 are 0. The group with all 0s is known as the reference group, which in this example is Group 4.

### Dummy Coding in ANOVA

The use of nominal data in prediction requires the use of dummy codes; this is because data need to be represented quantitatively for predictive purposes, and nominal data lack this quality. Once the data are coded properly, the analysis can be interpreted in a manner similar to traditional ANOVA.

Suppose we have three groups of people, single, married, and divorced, and we want to estimate their life satisfaction. In the following table, the first column identifies the single group (observations of single status are dummy coded as 1 and 0 otherwise ), and the second column identifies the married group (observations of married status are dummy coded as 1 and 0 otherwise). The divorced

group is left over, meaning this group is the reference group. However, the overall results will be the same no matter which groups we select.

| *Group* | *Satisfaction* | *Column 1* | *Column 2* | *Satis Group Mean* |
|---|---|---|---|---|
| Single | 25 | 1 | 0 | 24.80 |
| S | 28 | 1 | 0 | |
| S | 20 | 1 | 0 | |
| S | 26 | 1 | 0 | |
| S | 25 | 1 | 0 | |
| Married | 30 | 0 | 1 | 30.20 |
| M | 28 | 0 | 1 | |
| M | 32 | 0 | 1 | |
| M | 33 | 0 | 1 | |
| M | 28 | 0 | 1 | |
| Divorced | 20 | 0 | 0 | 23.8 |
| D | 22 | 0 | 0 | |
| D | 28 | 0 | 0 | |
| D | 25 | 0 | 0 | |
| D | 24 | 0 | 0 | |
| *Grand Mean* | *26.27* | *0.33* | *0.33* | |

Note there are three groups and thus 2 degrees of freedom between groups. Accordingly, there are two dummy-coded variables. If X1 denotes single, X2 denotes married, and X3 denotes divorced, then the single group is identified when X1 is 1 and X2 is 0; the married group is identified when X2 is 1 and X1 is 0; and the divorced group is identified when both X1 and X2 are 0.

If $\hat{Y}$ denotes the predicted level of life satisfaction, then we get the following regression equation:

$$\hat{Y} = a + b1(X1) + b2(X2),$$

where a is the interception, and b1 and b2 are slopes or weights. The divorced group is identified when both X1 and X2 are 0, so it drops out of the regression equation, leaving the predicted value equal to the mean of the divorced group.

The group that gets all 0s is the reference group. For this example, the reference group is the divorced group. The regression coefficients present a contrast or difference between the group identified by the column and the reference group. To be specific, the first b weight corresponds to the single group and the b1 represents the difference between the means of the divorced and single groups. The second b weight represents the difference in means between the divorced and married groups.

## Dummy Coding in Multiple Regression With Categorical Variables

Multiple regression is a linear transformation of the X variables such that the sum of squared deviations of the observed and predicted Y is minimized. The prediction of Y is accomplished by the following equation:

$$Y'_i = b_0 + b_1X_{1i} + b_2X_{2i} + \cdots + b_kX_{ki}$$

Categorical variables with two levels may be entered directly as predictor variables in a multiple regression model. Their use in multiple regression is a straightforward extension of their use in simple linear regression. When they are entered as predictor variables, interpretation of regression weights depends on how the variable is coded. When a researcher wishes to include a categorical variable with more than two levels in a multiple regression prediction model, additional steps are needed to ensure that the results are interpretable. These steps include recoding the categorical variable into a number of separate, dichotomous variables: dummy coding.

### Example Data: Faculty Salary Data

| *Faculty* | *Salary* | *Gender* | *Rank* | *Dept* | *Years* | *Merit* |
|---|---|---|---|---|---|---|
| 1 | Y1 | 0 | 3 | 1 | 0 | 1.47 |
| 2 | Y2 | 1 | 2 | 2 | 8 | 4.38 |
| 3 | Y3 | 1 | 3 | 2 | 9 | 3.65 |
| 4 | Y4 | 1 | 1 | 1 | 0 | 1.64 |
| 5 | Y5 | 1 | 1 | 3 | 0 | 2.54 |
| 6 | Y6 | 1 | 1 | 3 | 1 | 2.06 |
| 7 | Y7 | 0 | 3 | 1 | 4 | 4.76 |
| 8 | Y8 | 1 | 1 | 2 | 0 | 3.05 |
| 9 | Y9 | 0 | 3 | 3 | 3 | 2.73 |
| 10 | Y10 | 1 | 2 | 1 | 0 | 3.14 |

The simplest case of dummy coding is one in which the categorical variable has three levels and is converted to two dichotomous variables.

For example, *Dept* in the example data has three levels, 1 = Psychology, 2 = Curriculum, and 3 = Special Education. This variable could be dummy coded into two variables, one called *Psyc* and one called *Curri*. If Dept = 1, then Psyc would be coded with a 1 and Curri with a 0. If Dept = 2, then Psyc would be coded with a 0 and Curri would be coded with a 1. If Dept = 3, then both Psyc and Curri would be coded with a 0. The dummy coding is represented below.

| | | *Dummy Coded Variables* | |
|---|---|---|---|
| | *Dept* | *Psyc* | *Curri* |
| Psychology | 1 | 1 | 0 |
| Curriculum | 2 | 0 | 1 |
| Special Education | 3 | 0 | 0 |

A listing of the recoded data is presented below.

| *Faculty* | *Dept* | *Psyc* | *Curri* | *Salary* |
|---|---|---|---|---|
| 1 | 1 | *1* | *0* | Y1 |
| 2 | 2 | *0* | *1* | Y2 |
| 3 | 2 | *0* | *1* | Y3 |
| 4 | 1 | *1* | *0* | Y4 |
| 5 | 3 | *0* | *0* | Y5 |
| 6 | 3 | *0* | *0* | Y6 |
| 7 | 1 | *1* | *0* | Y7 |
| 8 | 2 | *0* | *1* | Y8 |
| 9 | 3 | *0* | *0* | Y9 |
| 10 | 1 | *1* | *0* | Y10 |

Suppose we get the following model summary table and coefficients table.

| *Model* | *R* | *R Square* | *Adjusted R Square* | *Std. Error of the Estimate* |
|---|---|---|---|---|
| 1 | .604 | .365 | .316 | 7.57 |

| *Change Statistics* | | | | |
|---|---|---|---|---|
| *R Square Change* | *F Change* | *df1* | *df2* | *Sig. F Change* |
| .365 | 7.472 | 2 | 26 | .003 |

| | | *Unstandardized Coefficients* | *Standardized Coefficients* | | |
|---|---|---|---|---|---|
| *Model* | | *B* | *Std. Error* | *Beta* | *t* | *Sig.* |
| 1 | (Constant) | 54.600 | 2.394 | | 22.807 | .000 |
| | Psyc | –8.886 | 3.731 | –.423 | –2.382 | .025 |
| | Curri | –12.350 | 3.241 | –.676 | –3.810 | .001 |

*Notes:* Coefficients. Dependent Variable: SALARY.
Psyc = Psychology; Curri = Curriculum.

The coefficients table can be interpreted as follows: The Psychology faculty makes $8,886 less in salary per year relative to the Special Education faculty, while the Curriculum faculty makes $12,350 less than the Special Education department.

## Combinations and Interaction of Categorical Predictor Variables

The previous examples dealt with individual categorical predictor variables with two or more levels. The following example illustrates how to create a new dummy coded variable that represents the interaction of certain variables.

Suppose we are looking at how gender, parental responsiveness, and the combination of gender and parental responsiveness influence children's social confidence. Confidence scores serve as the dependent variable, with gender and parental responsiveness scores (response) serving as the categorical independent variables. Response has three levels: high level, medium level, and low level. The analysis may be thought of as a two-factor ANOVA design, as below:

| | *Response Scale Values* | | | |
|---|---|---|---|---|
| *Gender* | *0 (Low)* | *1 (Mid)* | *2 (High)* | *Marginal Mean* |
| 0 (male) | $X_{11}$ | $X_{12}$ | $X_{13}$ | $M_{male}$ |
| 1 (female) | $X_{21}$ | $X_{22}$ | $X_{23}$ | $M_{female}$ |
| *Marginal Mean* | $M_{low}$ | $M_{mid}$ | $M_{high}$ | *Grand Mean* |

There are three "sources of variability" that could contribute to the differential explanation of confidence scores in this design. One source is the *main effect* for gender, another is the main effect for response, and the third is the *interaction effect*

between gender and response. Gender is already dummy coded in the data file with males = 0 and females = 1. Next, we will dummy code the response variable into two dummy coded variables, one new variable for each degree among groups for the response main effect (see below). Note that the low level of response is the reference group (shaded in the table below). Therefore, we have 2 (i.e., 3 − 1) dummy-coded variables for response.

| | *Dummy Coding* | | |
|---|---|---|---|
| *Original Coding* | *Low* | *Medium* | *High* |
| 0 (Low) | 0 | 0 | 0 |
| 1 (Medium) | 0 | 1 | 0 |
| 2 (High) | 0 | 0 | 1 |

Last comes the third dummy-coded variables, which represent the interaction source of variability. The new variables are labeled as *G*Rmid* (meaning gender interacting with the medium level of response), *G*Rhigh* (meaning gender interacting with the high level of response), and *G*Rlow* (meaning gender interacting with the low level of response), which is the reference group. To dummy code variables that represent interaction effects of categorical variables, we simply use the products of the dummy codes that were constructed separately for each of the variables. In this case, we simply multiply gender by dummy coded response. Note that there are as many new interaction dummy coded variables created as there are degrees of freedom for the interaction term in the ANOVAs design. The newly dummy coded variables are as follows:

| *Gender (main effect)* | *Response: Medium (main effect)* | *Response: High (main effect)* | *G*Rmid (interaction)* | *G*Rhigh (interaction)* |
|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 |
| 0 | 1 | 0 | 0 | 0 |
| 0 | 1 | 0 | 0 | 0 |
| 0 | 0 | 1 | 0 | 0 |
| 0 | 0 | 1 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 |
| 1 | 1 | 0 | 1 | 0 |
| 1 | 1 | 0 | 1 | 0 |
| 1 | 0 | 1 | 0 | 1 |
| 1 | 0 | 1 | 0 | 1 |

As in an ANOVA design analysis, the first hypothesis of interest to be tested is the interaction effect. In a multiple regression model, the first analysis tests this effect in terms of its "unique contribution" to the explanation of confidence scores. This can be realized by entering the response dummy-coded variables as a block into the model after gender has been entered. The Gender Response dummy-coded interaction variables are therefore entered as the last block of variables in creating the "full" regression model, that is, the model with all three effects in the equation. Part of the coefficients table is presented below.

| *Model* | | *B* | *t* | *p* |
|---|---|---|---|---|
| 1 | (Constant) | 18.407 | 29.66 | .000** |
| | Gender | 2.073 | 2.453 | .017* |
| 2 | (Constant) | 19.477 | 22.963 | .000** |
| | Gender | 2.045 | 2.427 | .018* |
| | Response (mid) | –1.401 | –1.518 | .133 |
| | Response (high) | –2.270 | –1.666 | .100 |
| 3 | (Constant) | 20.229 | 19.938 | .000** |
| | Gender | .598 | .425 | .672 |
| | Response (mid) | –1.920 | –1.449 | .152 |
| | Response (high) | –5.187 | –2.952 | .004* |
| | G*Rmid | 1.048 | .584 | .561 |
| | G*Rhigh | 6.861 | 2.570 | .012 |

*Notes:* $*p < .025$. $**p < .001$.

The regression equation is

$$\hat{Y} = 20.229 + .598X1 - 1.920X2 - 5.187X3 + 1.048X4 - 6.861X5,$$

in which X1 = gender (female), X2 = response (medium level), X3 = response (high level), X4 = Gender by Response (Female with Mid Level), and X5 = Gender by Response (Female with High Level). Now, b1 = .598 tells us girls receiving a low level of parental response have higher confidence scores than boys receiving the same level of parental response, but this difference is not significant ($p = .672 > .05$), and b2 = −1.920 tells us children receiving a medium level of parental response have lower confidence scores than children receiving a low level of parental response, but this difference is not significant ($p = .152 > .05$). However, $b_5 = 6.861$ tells us girls receiving a high

level of parental response tend to score higher than do boys receiving this or other levels of response, and this difference is significant ($p = .012 < .025$). The entire model tells us the pattern of mean confidence scores across the three parental response groups for boys is sufficiently different from the pattern of mean confidence scores for girls across the three parental response groups ($p < .001$).

*Jie Chen*

***See also*** Categorical Variable; Estimation

**Further Readings**


Aguinis, H. (2004). *Regression analysis for categorical moderators*. New York: Guilford.

Aiken, L. S., & West, S. G. (1991). *Multiple regression, testing and interpreting interaction*. Thousand Oaks, CA: Sage.

Allison, P. D. (1999). *Multiple regression: A primer*. Thousand Oaks, CA: Pine Forge Press.

Brannick, M. T. *Categorical IVs: Dummy, effect, and orthogonal coding*. Retrieved September 15, 2009, from http://luna.cas.usf.edu/~mbrannic/files/regression/anova1.html

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Gupta, R. (2008). *Coding categorical variables in regression models: Dummy and effect coding*. Retrieved September 15, 2009, from http://www.cscu.cornell.edu/news/statnews/stnews72.pdf

Keith, T. Z. (2006). *Multiple regression and beyond*. Boston: Allyn & Bacon.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction*. Eagan, MN: Thomson Learning.

Stockburger, D. W. (1997). Multiple regression with categorial variables. Retrieved September 15, 2009, from http://www.psychstat.missouristate.edu/multibook/mlt08m.html.

Warner, R. M. (2008). *Applied statistics: From bivariate through multivariate techniques*. Thousand Oaks, CA: Sage.


# Duncan's Multiple Range Test

*Duncan's multiple range test*, or *Duncan's test*, or *Duncan's new multiple range test,* provides significance levels for the difference between any pair of means, regardless of whether a significant *F* resulted from an initial analysis of variance. Duncan's test differs from the Newman–Keuls test (which slightly preceded it) in that it does not require an initial significant analysis of variance. It is a more powerful (in the statistical sense) alternative to almost all other post hoc methods.

When introducing the test in a 1955 article in the journal *Biometrics,* David B. Duncan described the procedures for identifying which pairs of means resulting from a group comparison study with more than two groups are significantly different from each other. Some sample mean values taken from the example presented by Duncan are given. Duncan worked in agronomics, so imagine that the means represent agricultural yields on some metric. The first step in the analysis is to sort the means in order from lowest to highest, as shown.

| *Groups* | *A* | *F* | *G* | *D* | *C* | *B* | *E* |
|---|---|---|---|---|---|---|---|
| Means | 49.6 | 58.1 | 61.0 | 61.5 | 67.6 | 71.2 | 71.3 |

From tables of values that Duncan developed from the *t*-test formula, standard critical differentials at the .05 level are identified. These are significant *studentized* differences, which must be met or surpassed. To maintain the nominal significance level one has chosen, these differentials get slightly higher as the two means that are compared become further apart in terms of their rank ordering. In the example shown, the means for groups A and F have an interval of 2 because they are adjacent to each other. Means A and E have an interval of 7 as there are seven means in the span between them. By multiplying the critical differentials by the standard error of the mean, one can compute the *shortest significant ranges* for each interval width (in the example, the possible intervals are 2, 3, 4, 5, 6, and 7). With the standard error of the mean of 3.643 (which is supplied by Duncan for this example), the *shortest significant ranges* are calculated.

| *Range* | *2* | *3* | *4* | *5* | *6* | *7* |
|---|---|---|---|---|---|---|
| Studentized differences | 2.89 | 3.04 | 3.12 | 3.20 | 3.25 | 3.29 |
| Shortest significant ranges | 10.53 | 11.07 | 11.37 | 11.66 | 11.84 | 11.99 |

For any two means to be significantly different, their distance must be equal to or greater than the associated shortest significant range. For example, the distance between mean F (58.1) and mean B (71.2) is 13.1. Within the rank ordering of the means, the two means form an interval of width 5, with an associated shortest significant range of 11.66. Because $13.1 > 11.66$, the two means are significantly different at the .05 level.

Duncan suggested a graphical method of displaying all possible mean comparisons and whether they are significant compared with one another. This method involved underlining those clusters of means that are not statistically different. Following his suggestion, the results for this sample are shown below.

| *Groups* | *A* | *F* | *G* | *D* | *C* | *B* | *E* |
|---|---|---|---|---|---|---|---|
| Means | 49.6 | 58.1 | 61.0 | 61.5 | 67.6 | 71.2 | 71.3 |

*Notes:* Means that are underlined by the same line are not significantly different from each other. Means that are underlined by different lines are significantly different from each other.

| | *Number of Means* | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *Degrees of Freedom* | *2* | *3* | *4* | *5* | *6* | *7* | *8* | *9* | *10* |
| *5* | 3.64 | 3.74 | 3.79 | 3.83 | 3.83 | 3.83 | 3.83 | 3.83 | 3.83 |
| 10 | 3.15 | 3.30 | 3.37 | 3.43 | 3.46 | 3.47 | 3.47 | 3.47 | 3.47 |
| 15 | 3.01 | 3.16 | 3.25 | 3.31 | 3.36 | 3.38 | 3.40 | 3.42 | 3.43 |
| 20 | 2.95 | 3.10 | 3.18 | 3.25 | 3.30 | 3.34 | 3.36 | 3.38 | 3.40 |
| 30 | 2.89 | 3.04 | 3.12 | 3.20 | 3.25 | 3.29 | 3.32 | 3.35 | 3.37 |
| 60 | 2.83 | 2.98 | 3.08 | 3.14 | 3.20 | 3.24 | 3.28 | 3.31 | 3.33 |
| 100 | 2.80 | 2.95 | 3.05 | 3.12 | 3.18 | 3.22 | 3.26 | 3.29 | 3.32 |

*Note:* Significant Studentized Ranges at the .05 Level for Duncan's Multiple Range Test.

The philosophical approach taken by Duncan is an unusually liberal one. It allows for multiple pairwise comparisons without concern for inflation of the Type I error rate. A researcher may perform dozens of post hoc analyses in the absence of specific hypotheses and treat all tests as if they are conducted at the .05 (or whatever the nominal value chosen) level of significance. The comparisons may be analyzed even in the absence of an overall *F* test indicating that any differences exist. Not surprisingly, Duncan's multiple range test is not recommended by many statisticians who prefer more conservative approaches that minimize the Type I error rate. Duncan's response to those concerns was to argue that because the null hypothesis is almost always known to be false to begin with, it is more reasonable to be concerned about making Type II errors, missing true population differences, and his method certainly minimizes the true Type II error rate.

A small table that includes the significant studentized differences calculated by Duncan and reported in his 1955 paper is provided below. It shows values based on the number of means to be considered and the degrees of freedom in the experiment. The table assumes a .05 alpha level.

*Bruce Frey*

***See also*** Newman–Keuls Test and Tukey Test; Post Hoc Comparisons; Type II Error

**Further Readings**

Chew, V. (1976). Uses and abuses of Duncan's Multiple Range Test. *Proceedings of the Florida State Horticultural Society, 89,* 251–253.

Duncan, D. B. (1955). Multiple range and multiple *F* tests. *Biometrics, March,* 1–42.

# Dunnett's Test

Dunnett's test is one of a number of a posteriori or post hoc tests, run after a significant one-way analysis of variance (ANOVA), to determine which differences are significant. The procedure was introduced by Charles W. Dunnett in 1955. It differs from other post hoc tests, such as the Newman–Keuls test, Duncan's Multiple Range test, Scheffé's test, or Tukey's Honestly Significant Difference test, in that its use is restricted to

comparing a number of experimental groups against a single control group; it does not test the experimental groups against one another. Background information, the process of running Dunnett's test, and an example are provided in this entry.

## Background

A one-way ANOVA tests the null hypothesis ($H_0$) that all the $k$ treatment means are equal; that is,

$$H_0 : \mu_A = \mu_B = \mu_C = \cdots = \mu_k,$$

against the alternative hypothesis ($H_1$) that at least one of the means is different from the others. The difficulty is that if $H_0$ is rejected, it is not known which mean differs from the others. It is possible to run $t$ tests on all possible pairs of means (e.g., A vs. B, A vs. C, B vs. C). However, if there are five groups, this would result in 10 $t$ tests (in general, there are $k \times (k-1) / 2$ pairs). Moreover, the tests are not independent, because any one mean enters into a number of comparisons, and there is a common estimate of the experimental error. As a result, the probability of a Type I error (that is, concluding that there is a significant difference when in fact there is not one) increases beyond 5% to an unknown extent. The various post hoc tests are attempts to control this *family-wise error rate* and constrain it to 5%.

The majority of the post hoc tests compare each group mean against every other group mean. One, called Scheffé's test, goes further, and allows the user to compare combinations of groups (e.g., A + B vs. C, A + C vs. B, A + B vs. C + D). Dunnett's test is limited to the situation in which one group is a control or reference condition ($R$), and each of the other (experimental group) means is compared to it.

## Dunnett's Test

Rather than one null hypothesis, Dunnett's test has as many null hypotheses as there are experimental groups. If there are three such groups (A, B, and C), then the null and alternative hypotheses are:

$$H_{0A} : \mu_A = \mu_R;\ H_{1A} : \mu_A \neq \mu_R$$

$$H_{0B} : \mu_B = \mu_R;\ H_{1B} : \mu_B \neq \mu_R$$

$$H_{0C} : \mu_C = \mu_R;\ H_{1C} : \mu_C \neq \mu_R$$

Each hypothesis is tested against a critical value, called $q'$, using the formula:

$$q' = \frac{\overline{X}_i - \overline{X}_R}{\sqrt{2 \times MS_{\text{error}} \left(\frac{1}{n_i} + \frac{1}{n_R}\right)}},$$

where the subscript $i$ refers to each of the experimental groups, and $MS_{\text{error}}$ is the mean square for the error term, taken from the ANOVA table. The value of $q'$ is then compared against a critical value in a special table, the researcher knowing the number of groups (including the reference group) and the degrees of freedom of $MS_{\text{error}}$. The form of the equation is very similar to that of the Newman–Keuls test. However, it accounts for the fact that there are a smaller number of comparisons because the various experimental groups are not compared with each other. Consequently, it is more powerful than other post hoc tests.

In recent years, the American Psychological Association strongly recommended reporting confidence intervals (CIs) in addition to point estimates of a parameter and $p$ levels. It is possible to both derive CIs and run Dunnett's test directly, rather than first calculating individual values of $q'$ for each group. The first step is to determine an allowance ($A$), defined as

$$A = t_{1-(\alpha/2)} \sqrt{\frac{2 \times MS_{\text{error}}}{n_h}},$$

where $t_{1-(\alpha/2)}$ is taken from the Dunnett tables, and $n_h$ is the harmonic mean of the sample sizes of all the groups (including the reference), which is

$$n_h = \frac{k}{\frac{1}{n_1} + \frac{1}{n_2} + \frac{1}{n_3} + \cdots + \frac{1}{n_k}}.$$

Then the CI for each term is

$$(\overline{X}_i - \overline{X}_R) \pm A$$

The difference is statistically significant if the CI does not include zero.

**Table 1** Results of a Fictitious Experiment Comparing Three Treatments Against a Reference Condition

| *Group* | *Mean* | *Standard Deviation* |
|---|---|---|
| Reference | 50.00 | 3.55 |
| A | 61.00 | 4.24 |
| B | 52.00 | 2.45 |
| C | 45.00 | 3.74 |

## An Example

Three different drugs (A, B, and C) are compared against placebo (the reference condition, R), and there are 4 participants per group. The results are shown in Table 1, and the ANOVA summary in Table 2. For Group A, $q'$ is:

$$q' = \frac{61 - 50}{\sqrt{2 \times 12.667\left(\frac{1}{4} + \frac{1}{4}\right)}} = \frac{11}{\sqrt{12.667}} = 3.09$$

For Group B, $q' = 0.56$, and it is 1.40 for Group C. With four groups and $df_{\text{error}} = 12$, the critical value in Dunnett's table is 2.68 for $\alpha = .05$ and 3.58 for $\alpha = .01$. We therefore conclude that Group A is different from the reference group at $p < .05$, and that Groups B and C do not differ significantly from it.

The value of $A$ is :

$$A = 2.68\sqrt{\frac{2 \times 12.667}{4}} = 2.68 \times 2.52 = 6.75,$$

**Table 2** The ANOVA Table for the Results in Table 1

| *Source of Variance* | *Sum of Squares* | *df* | *Mean Square* | *F* | *p* |
|---|---|---|---|---|---|
| Between groups | 536.000 | 3 | 178.667 | 14.105 | $< .001$ |
| Within groups | 152.000 | 12 | 12.667 | | |
| Total | 688.000 | 15 | | | |

meaning that the 95% CI for Group A is

$$(61 - 50) \pm 6.75 = 4.25 - 17.75.$$

*David L. Streiner*

***See also*** Analysis of Variance (ANOVA); Duncan's Multiple Range Test; Honestly Significant Difference (HSD) Test; Multiple Comparison Tests; Newman–Keuls Test and Tukey Test; Pairwise Comparisons; Scheffé Test; Tukey's Honestly Significant Difference (HSD)

### Further Readings

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association, 50,* 1096–1121.

Dunnett, C. W. (1964). New tables for multiple comparisons with a control. *Biometrics, 20,* 482–491.

Sahai, H., & Ageel, M. I. (2000). *Analysis of variance: Fixed, random, and mixed models.* New York: Birkhäuser.

# E

## Ecological Validity

Ecological validity is the degree to which test performance predicts behaviors in real-world settings. Today, psychologists are called upon by attorneys, insurance agencies, vocational rehabilitation counselors, and employers to draw inferences about clients' cognitive capacities and their implications in real-world settings from psychological tests. These demands have accentuated the importance of ecological validity. Originally, neuropsychological tests were created as tools for detecting and localizing neuropathology. The diagnostic utility of such assessment instruments decreased with the development of brain-imaging techniques, and neuropsychology shifted its focus toward identifying the practical implications of brain pathology. Society's increasing interest in clients' everyday abilities has necessitated further research into the ecological validity of psychological and neuropsychological tests. The dimensions, applications, limitations, and implications of ecological validity are discussed in this entry.

### Dimensions

Robert Sbordone refers to ecological validity as "the functional and predictive relationship between the patient's performance on a set of neuropsychological tests and the patient's behavior in a variety of real-world settings" (Sbordone & Long, p. 16). This relationship is not absolute; tests tend to fall on a continuum ranging from low to high levels of ecological validity. Despite no universally agreed-upon definition of ecological validity, a deeper understanding of the concept can be achieved by analyzing its three dimensions: test environment, stimuli under examination, and behavioral response.

#### *Test Environment*

In the field of psychological assessment, controlled test environments are recommended to allow the client's "best performance," and psychologists have attempted to reduce distractions, confusion, and fatigue in the testing situation. Historically, to avoid misdiagnosing brain pathology, evaluating a client's best performance was crucial. However, because neuropsychologists today are asked to predict clients' functioning in real-world settings, the ecological validity of the traditional test environment has been called into question. Unlike testing situations, the natural world does not typically provide a quiet, supportive, distraction-reduced environment. The disparity between test environments and clients' everyday environments may reduce the predictive accuracy of psychological assessments. Today, many referral questions posed to neuropsychologists call for the development of testing environments that more closely approximate real-world settings.

#### *Stimuli Under Examination*

The extent to which stimuli used during testing resemble stimuli encountered in daily life should

be taken into account when evaluating ecological validity. For example, the Grocery List Selective Reminding Test is a test that uses real-world stimuli. Unlike traditional paired associate or list-learning tests, which often use arbitrary stimuli, the Grocery List Selective Reminding Test employs a grocery list to evaluate verbal learning. Naturally occurring stimuli increase the ecological validity of neuropsychological tests.

### *Behavioral Response*

Another important dimension of ecological validity is assuring that behavioral responses elicited are representative of the person's natural behaviors and appropriately related to the construct being measured. Increased levels of ecological validity would be represented in simulator assessment of driving by moving the cursor with the arrow keys, with the mouse, or with a steering wheel. The more the response approximates the criterion, the greater the ecological validity.

The two main methods of establishing ecological validity are veridicality and verisimilitude. These methods are related to, but not isomorphic with, the traditional constructs of concurrent validity/predictive validity and construct validity/face validity, respectively.

#### Veridicality

Veridicality is the degree to which test scores correlate with measures of real-world functioning. The veridicality approach examines the statistical relationship between performance on traditional neuropsychological tests and one or more selected outcome measures, including self-reports, informant questionnaires, clinician ratings, performance-based measures, employment status, and activities of daily living. Self-reports have repeatedly been shown as weaker predictors of everyday performance than clinician and informant ratings. However, with recent advances in technology, researchers have attempted to increase the ecological validity of self-reports by using ambulatory monitoring devices to conduct ecological momentary assessments to measure patients' behaviors, moods, perception of others, physiological variables, and physical activities in natural settings. This technology is in its infancy, and more research is needed to determine its effectiveness and utility. Commonly used outcome measures to which traditional neuropsychological tests are correlated in the veridicality approach are the Dysexecutive Questionnaire (DEX) and the Behavior Rating Inventory of Executive Functioning.

One limitation of the veridicality approach is that the outcome measures selected for comparison with the traditional neuropsychological test may not accurately represent the client's everyday functioning. Also, many of the traditional neuropsychological tests evaluated using the veridicality approach were developed to diagnose brain pathology, not make predictions about daily functioning.

#### Verisimilitude

Verisimilitude is the degree to which tasks performed during testing resemble tasks performed in daily life. With the verisimilitude approach, tests are created to simulate real-world tasks. Some limitations of the verisimilitude approach include the cost of creating new tests and the reluctance of clinicians to put these new tests into practice. Mere face validity cannot be substituted for empirical research when assessing the ecological validity of neuropsychological tests formed from this approach.

## Ecological Validity of Neuropsychological Tests

### *Executive Functioning*

Although there are data to support the ecological validity of traditional neuropsychological tests of executive functioning, a growing body of literature suggests that the traditional tests (i.e., Wisconsin Card Sorting Test, Stroop Color-Word Test, Trail Making Test, and Controlled Oral Word Association Test), at best, only moderately predict everyday executive functioning and that tests developed with ecological validity in mind are more effective. Studies examining the relationship between the Hayling and Brixton tests and the DEX have reported favorable results in patients with nondegenerative brain disorders, frontal lobe lesions, and structural brain damage. The California Verbal Learning Test has been used effectively to predict job performance and occupational status, whereas the preservative responses of the

Wisconsin Card Sorting Test are capable of effectively predicting occupational status only. Newer tests are being developed to encompass verisimilitude in the study of executive functioning. These tests include the Virtual Planning Test and the Behavioral Assessment of Dysexecutive Syndrome.

### *Attention*

Although research on the veridicality of tests of attention is limited, there is reasonable evidence that traditional tests of attention are ecologically valid. More research should be conducted to verify current results, but the ecological validity of these traditional tests is promising. Although some investigators are not satisfied with traditional tests for attention deficit/hyperactivity disorder (ADHD), researchers have found evidence of predictive validity in the Hayling test in children with ADHD. The Test of Everyday Attention (TEA), which was developed using the verisimilitude approach, is an assessment tool designed to evaluate attentional switching, selective attention, and sustained attention. Investigators have found correlations between the TEA and other standardized measures of attention, including the Stroop Color-Word Test, the Symbol Digit Modalities Test, and the Paced Auditory Serial Addition Tests.

### *Memory Tests*

The Rivermead Behavioral Memory Test (RBMT), designed using the verisimilitude approach, is a standardized test used to assess everyday memory functioning. The memory tasks in the RBMT resemble everyday memory demands, such as remembering a name or an appointment. Significant correlations have been demonstrated between the RBMT and other traditional memory tests as well as between the RBMT and ratings of daily functioning by subjects, significant others, and clinicians. Some studies have revealed the superiority of the RBMT and the TEA in predicting everyday memory functioning or more general functioning when compared to more traditional neuropsychological tests. In addition to the RBMT, other tests that take verisimilitude into account include the 3-Objects-3-Places, the Process Dissociation Procedure, and the Memory in Reality. Research also suggests that list learning tasks have suitable ecological validity to aid the diagnosis and outcome prediction of patients with epilepsy.

### *Perception*

Research on ecological validity of perceptual tests is limited. The Behavioral Inattention Test was developed to assist the prediction of everyday problems arising from unilateral visual neglect. Ecological validity of the Wechsler Adult Intelligence Scale—Revised has been shown for assessing visuoconstructive skills. Investigators used subtests such as Block Design, Object Assembly, and Picture Completion and found that poor performance predicts problems in daily living.

### *Virtual Tests*

With increasing advances in cyber technology, neuropsychological assessments are turning to computers as an alternative to real-world behavioral observations. One innovative approach has been the use of virtual reality scenarios where subjects are exposed to machines that encompass 3-D, real-world-like scenes and are asked to perform common functions in these environments allowing naturalistic stimulus challenges while maintaining experimental control. These tests include the Virtual Reality Cognitive Performance Assessment Test, a virtual city, the Virtual Office, and a simulated street to assess memory and executive functioning. These methods suggest that the virtual tests may provide a new, ecological measure for examining memory deficits in patients.

## Other Applications

### *Academic Tests*

There is a high degree of variance in our educational systems, from grading scales and curricula to expectations and teacher qualifications. Because of such high variability, colleges and graduate programs use standardized tests as part of their admissions procedure. Investigators have found that the American College Test has low predictive validity of first-year grades as well as graduation grades for students attending undergraduate programs. Also, correlations have been found between Scholastic Assessment Test Math (SATM) and Scholastic Assessment Test Verbal (SATV) test scores and

overall undergraduate GPA, but the SATM may underpredict women's grades.

Studies suggest that the Graduate Record Exam (GRE) is capable of at least modestly predicting first-year grades in graduate school and veterinary programs as well as graduate grade point average, faculty ratings, comprehensive examination scores, citation counts and degree attainment across departments, acceptance into PhD programs, external awards, graduation on time, and thesis publication. Although some studies suggest that the GRE is an ecologically valid tool, there is debate about how much emphasis to place on GRE scores in the postgraduate college admissions process. The Medical College Admission Test (MCAT) has been shown to be predictive of success on written tests assessing skills in clinical medicine. In addition, the MCAT was able to positively predict performance on physician certification exams.

### *Activities of Daily Living*

To evaluate patients' ability to function independently, researchers have investigated the accuracy of neuropsychological tests in predicting patients' capacities to perform activities of daily living (ADL), such as walking, bathing, dressing, and eating. Research studies found that neuropsychological tests correlated significantly with cognitive ADL skills involving attention and executive functioning. Overall, ADL research demonstrates low to moderate levels of ecological validity. Ecological validity is improved, however, when the ADLs evaluated have stronger cognitive components. Driving is an activity of daily living that has been specifically addressed in ecological validity literature, and numerous psychometric predictors have been identified. But none does better at prediction than the actual driving of a small-scale vehicle on a closed course. In like manner, a wheelchair obstacle course exemplifies an ecologically valid outcome measure for examining the outcome of visual scanning training in persons with right brain damage.

### *Vocational Rehabilitation*

Referral questions posed to neuropsychologists have shifted from diagnostic issues to rehabilitative concerns. In an effort to increase the ecological validity of rehabilitation programs, the disability management movement, which uses work environments as rehabilitation sites instead of vocational rehabilitation centers, emerged.

The Behavioral Assessment of Vocational Skills is a performance-based measure able to significantly predict workplace performance. Also, studies have shown that psychosocial variables significantly predict a patient's ability to function effectively at work. When predicting employment status, the Minnesota Multiphasic Personality Inventory is one measure that has been shown to add ecological validity to neuropsychological test performance.

### *Employment*

Prediction of a person's ability to resume employment after disease or injury has become increasingly important as potential employers turn to neuropsychologists with questions about job capabilities, skills, and performance. Ecological validity is imperative in assessment of employability because of the severe consequences of inaccurate diagnosis. One promising area of test development is simulated vocational evaluations (SEvals), which ask participants to perform a variety of simulated vocational tasks in environments that approximate actual work settings. Research suggests that the SEval may aid evaluators in making vocational decisions. Other attempts at improving employment predictions include the Occupational Abilities and Performance Scale and two self-report questionnaires, the Work Adjustment Inventory and the Working Inventory.

### *Forensic Psychology*

Forensic psychology encompasses a vast spectrum of legal issues including prediction of recidivism, identification of malingering, and assessment of damages in personal injury and medico-legal cases. There is little room for error in these predictions as much may be at stake.

Multiple measures have been studied for the prediction of violent behavior, including (a) the Psychopathic checklist, which has shown predictive ability in rating antisocial behaviors such as criminal violence, recidivism, and response to correctional treatment; (b) the MMPI-2, which has a psychopathy scale (scale 4) and is sensitive to

antisocial behavior; (c) the California Psychological Inventory, which is a self-report questionnaire that provides an estimate of compliance with society's norms; and (d) the Mental Status Examination, where the evaluator obtains personal history information, reactions, behaviors, and thought processes. In juveniles, the Youth Level of Service Case Management Inventory (YLS/CMI) has provided significant information for predicting recidivism in young offenders; however, the percentage variance predicted by the YLS/CMI was low.

Neuropsychologists are often asked to determine a client's degree of cognitive impairment after a head injury so that the estimated lifetime impact can be calculated. In this respect, clinicians must be able to make accurate predictions about the severity of the cognitive deficits caused by the injury.

## Limitations and Implications for the Future

In addition to cognitive capacity, other variables that influence individuals' everyday functioning include environmental cognitive demands, compensatory strategies, and noncognitive factors. These variables hinder researchers' attempts at demonstrating ecological validity. With regard to environmental cognitive demands, for example, an individual in a more demanding environment will demonstrate more functional deficits in reality than an individual with the same cognitive capacity in a less demanding environment. To improve ecological validity, the demand characteristics of an individual's environment should be assessed. Clients' consistency in their use of compensatory strategies across situations will also affect ecological validity. Clinicians may underestimate a client's everyday functional abilities if compensatory strategies are not permitted during testing or if the client simply chooses not to use his or her typical repertoire of compensatory skills during testing. Also, noncognitive factors, including psychopathology, malingering, and premorbid functioning, impede the predictive ability of assessment instruments.

A dearth of standardized outcome measures, variable test selection, and population effects are other limitations of ecological validity research. Mixed results in current ecological validity literature may be a result of using inappropriate outcome measures. More directed hypotheses attempting to delineate the relationship between particular cognitive constructs and more specific everyday abilities involving those constructs may increase the ecological validity of neuropsychological tests. However, there is some disagreement as to which tests appropriately measure various cognitive constructs. Presently, comparing across ecological validity research studies is challenging because of the wide variety of outcome measures, neuropsychological tests, and populations assessed.

Great strides have been made in understanding the utility of traditional tests and developing new and improved tests that increase psychologists' abilities to predict people's functioning in everyday life. As our understanding of ecological validity increases, future research should involve more encompassing models, which take other variables into account aside from test results. Interviews with the client's friends and family, medical and employment records, academic reports, client complaints, and direct observations of the client can be helpful to clinicians faced with ecological questions. In addition, ecological validity research should address test environments, environmental demands, compensatory strategies, noncognitive factors, test and outcome measure selection, and population effects in order to provide a foundation from which general conclusions can be drawn.

*William Drew Gouvier, Alyse A. Barker, and Mandi Wilkes Musso*

***See also*** Concurrent Validity; Construct Validity; Face Validity; Predictive Validity

### Further Readings

Chaytor, N., & Schmitter-Edgecombe, M. (2003). The ecological validity of neuropsychological tests: A review of the literature on everyday cognitive skills. *Neuropsychology Review, 13*(4), 181–197.

Farias, S. T., Harrell, E., Neumann, C., & Houtz, A. (2003). The relationship between neuropsychological performance and daily functioning in individuals with Alzheimer's disease: Ecological validity of neuropsychological tests. *Archives of Clinical Neuropsychology, 18,* 655–672.

Farmer, J., & Eakman, A. (1995). The relationship between neuropsychological functioning and instrumental activities of daily living following acquired brain injury. *Applied Neuropsychology, 2,* 107–115.

Heaton, R. K., & Pendleton, M. G. (1981). Use of neuropsychological tests to predict adult patients' everyday functioning. *Journal of Consulting and Clinical Psychology, 49*(6), 807–821.
Kibby, M. Y., Schmitter-Edgecombe, M., & Long, C. J. (1998). Ecological validity of neuropsychological tests: Focus on the California Verbal Learning Test and the Wisconsin Card Sorting Test. *Archives of Clinical Neuropsychology, 13*(6), 523–534.
McCue, M., Rogers, J., & Goldstein, G. (1990). Relationships between neuropsychological and functional assessment in elderly neuropsychiatric patients. *Rehabilitation Psychology, 35*, 91–99.
Norris, G., & Tate, R. L. (2000). The Behavioural Assessment of the Dysexecutive Syndrome (BADS): Ecological concurrent and construct validity. *Neuropsychological Rehabilitation, 10*(1), 33–45.
Schmuckler, M. A. (2001). What is ecological validity? A dimensional analysis. *Infancy, 2*(4), 419–436.
Silver, C. H. (2000). Ecological validity of neuropsychological assessment in childhood traumatic brain injury. *Journal of Head Trauma Rehabilitation, 15*(4), 973–988.
Sbordone, R. J., & Long, C. (1996). *Ecological validity of neuropsychological testing*. Delray Beach, FL: GR Press/St. Lucie Press.
Wood, R. L., & Liossi, C. (2006). The ecological validity of executive tests in a severely brain injured sample. *Archives of Clinical Neuropsychology, 21*, 429–437.

# Effect Coding

Effect coding is a coding scheme used when an analysis of variance (ANOVA) is performed with multiple linear regression (MLR). With effect coding, the experimental effect is analyzed as a set of (nonorthogonal) contrasts that opposes all but one experimental condition to one given experimental condition (usually the last one). With effect coding, the intercept is equal to the grand mean, and the slope for a contrast expresses the difference between a group and the grand mean.

## Multiple Regression Framework

In linear multiple regression analysis, the goal is to predict, knowing the measurements collected on $N$ subjects, a dependent variable $Y$ from a set of $J$ independent variables denoted

$$\{X_1, \ldots, X_j, \ldots, X_J\}. \tag{1}$$

We denote by $\mathbf{X}$ the $N \times (J+1)$ augmented matrix collecting the data for the independent variables (this matrix is called augmented because the first column is composed only of 1s), and by $\mathbf{y}$ the $N \times 1$ vector of observations for the dependent variable. These two matrices have the following structure.

$$\mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,k} & \cdots & x_{1,K} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,k} & \cdots & x_{n,K} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 1 & x_{N,1} & \cdots & x_{N,k} & \cdots & x_{N,k} \end{bmatrix} \tag{2}$$

$$\text{and } \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \\ \vdots \\ y_N \end{bmatrix}.$$

The predicted values of the dependent variable $\hat{\mathbf{Y}}$ are collected in a vector denoted $\hat{\mathbf{y}}$ and are obtained as

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} \quad \text{with} \quad \mathbf{b} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}. \tag{3}$$

where $\mathbf{T}$ denotes the transpose of a matrix and the vector $\mathbf{b}$ has $J$ components. Its first component is traditionally denoted $\mathbf{b}_0$, it is called the intercept of the regression, and it represents the regression component associated with the first column of the matrix $\mathbf{X}$. The additional $J$ components are called *slopes*, and each of them provides the amount of change in $Y$ consecutive to an increase in one unit of its corresponding column.

The regression sum of squares is obtained as

$$SS_{\text{regression}} = \mathbf{b}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} - \frac{1}{N}(\mathbf{1}^{\mathrm{T}}\mathbf{y})^2 \tag{4}$$

(with $1^{\mathrm{T}}$ being a row vector of 1s conformable with $\mathbf{y}$). The total sum of squares is obtained as

$$SS_{\text{total}} = \mathbf{y}^{\mathrm{T}}\mathbf{y} - \frac{1}{N}(\mathbf{1}^{\mathrm{T}}\mathbf{y})^2. \tag{5}$$

The residual (or error) sum of squares is obtained as

$$SS_{\text{error}} = \mathbf{y}^{\mathrm{T}}\mathbf{y} - \mathbf{b}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y}. \quad (6)$$

The quality of the prediction is evaluated by computing the multiple coefficient of correlation denoted $R^2_{Y.1,\ldots,J}$. This coefficient is equal to the squared coefficient of correlation between the dependent variable ($Y$) and the predicted dependent variable ($\hat{Y}$).

An alternative way of computing the multiple coefficient of correlation is to divide the regression sum of squares by the total sum of squares. This shows that $R^2_{Y.1,\ldots,J}$ can also be interpreted as the proportion of variance of the dependent variable explained by the independent variables. With this interpretation, the multiple coefficient of correlation is computed as

$$R^2_{Y.1,\ldots,J} = \frac{SS_{\text{regression}}}{SS_{\text{regression}} + SS_{\text{error}}} = \frac{SS_{\text{regression}}}{SS_{\text{total}}}. \quad (7)$$

## Significance Test

In order to assess the significance of a given $R^2_{Y.1,\ldots,J}$, we can compute an $F$ ratio as

$$F = \frac{R^2_{Y.1,\ldots,J}}{1 - R^2_{Y.1,\ldots,J}} \times \frac{N - J - 1}{J}. \quad (8)$$

Under the usual assumptions of normality of the error and of independence of the error and the scores, this $F$ ratio is distributed under the null hypothesis as a Fisher distribution with $\nu_1 = J$ and $\nu_2 = N - J - 1$ degrees of freedom.

## Analysis of Variance Framework

For an ANOVA, the goal is to compare the means of several groups and to assess whether these means are statistically different. For the sake of simplicity, we assume that each experimental group comprises the same number of observations denoted I (i.e., we are analyzing a "balanced design"). So, if we have $K$ experimental groups with a total of $I$ observations per group, we have a total of $K \times I = N$ observations denoted $Y_{i,k}$. The first step is to compute the $K$ experimental means denoted $M_{+,k}$ and the grand mean denoted $M_{+,+}$. The ANOVA evaluates the difference between the mean by comparing the dispersion of the experimental means to the grand mean (i.e., the dispersion between means) with the dispersion of the experimental scores to the means (i.e., the dispersion within the groups). Specifically, the dispersion between the means is evaluated by computing the sum of squares between means, denoted $SS_{\text{between}}$, and computed as

$$SS_{\text{between}} = I \times \sum_{k}^{K} (M_{+,k} - M_{+,+})^2. \quad (9)$$

The dispersion within the groups is evaluated by computing the sum of squares within groups, denoted $SS_{\text{within}}$ and computed as

$$SS_{\text{within}} = \sum_{k}^{K} \sum_{i}^{I} (Y_{i,k} - M_{+,k})^2. \quad (10)$$

If the dispersion of the means around the grand mean is due only to random fluctuations, then the $SS_{\text{between}}$ and the $SS_{\text{within}}$ should be commensurable. Specifically, the null hypothesis of no effect can be evaluated with an $F$ ratio computed as

$$F = \frac{SS_{\text{between}}}{SS_{\text{within}}} \times \frac{N - K}{K - 1}. \quad (11)$$

Under the usual assumptions of normality of the error and of independence of the error and the scores, this $F$ ratio is distributed under the null hypothesis as a Fisher distribution with $\nu_1 = K - 1$ and $\nu_2 = N - K$ degrees of freedom. If we denote by $R^2_{\text{experimental}}$ the following ratio

$$R^2_{\text{experimental}} = \frac{SS_{\text{between}}}{SS_{\text{between}} + SS_{\text{within}}}, \quad (12)$$

we can re-express Equation 11 in order to show its similarity with Equation 8 as

$$F = \frac{R^2_{\text{experimental}}}{1 - R^2_{\text{experimental}}} \times \frac{N - K}{K - 1}. \quad (13)$$

## Analysis of Variance With Effect Coding Multiple Linear Regression

The similarity between Equations 8 for MLR and 13 for ANOVA suggests that these two methods are related, and this is indeed the case. In fact, the computations for an ANOVA can be performed with MLR via a judicious choice of

the matrix $\mathbf{X}$ (the dependent variable is represented by the vector $\mathbf{y}$). In all cases, the first column of $\mathbf{X}$ will be filled with 1s and is coding for the value of the intercept. One possible choice for $\mathbf{X}$, called *mean coding*, is to have one additional column in which the value for the $n$th observation will be the mean of its group. This approach provides a correct value for the sums of squares but not for the $F$ (which needs to be divided by $K-1$). Most coding schemes will use $J=K-1$ linearly independent columns (as many columns as there are degrees of freedom for the experimental sum of squares). They all give the same correct values for the sums of squares and the $F$ test but differ for the values of the intercept and the slopes. To implement effect coding, the first step is to select a group called the contrasting group; often, this group is the last one. Then, each of the remaining $J$ groups is contrasted with the contrasting group. This is implmented by creating a vector for which all elements of the contrasting group have the value $-1$, all elements of the group under consideration have the value of $+1$, and all other elements have a value of 0.

With the effect coding scheme, the intercept is equal to the grand mean, and each slope coefficient is equal to the difference between the grand mean and the mean of the group whose elements were coded with values of 1. This difference estimates the experimental effect of this group, hence the name of *effect* coding for this coding scheme. The mean of the contrasting group is equal to the intercept minus the sum of all the slopes.

## Example

The data used to illustrate effect coding are shown in Table 1. A standard ANOVA would give the results displayed in Table 2.

**Table 1** A Data Set for an ANOVA

| | $a_1$ | $a_2$ | $a_3$ | $a_4$ | |
|---|---|---|---|---|---|
| $S_1$ | 20 | 21 | 17 | 8 | |
| $S_2$ | 17 | 16 | 16 | 11 | |
| $S_3$ | 17 | 14 | 15 | 8 | |
| $M_{a.}$ | 18 | 17 | 16 | 9 | $M.. = 15$ |

*Note:* A total of $N = 12$ observations coming from $K = 4$ groups with $I = 3$ observations per group.

**Table 2** ANOVA Table for the Data From Table 1

| *Source* | *df* | *SS* | *MS* | *F* | *Pr(F)* |
|---|---|---|---|---|---|
| Experimental | 3 | 150.00 | 50.00 | 10.00 | .0044 |
| Error | 8 | 40.00 | 5.00 | | |
| Total | 11 | 190.00 | | | |

In order to perform an MLR analysis, the data from Table 1 need to be "vectorized" in order to provide the following $\mathbf{y}$ vector:

$$\mathbf{y} = \begin{bmatrix} 20 \\ 17 \\ 17 \\ 21 \\ 16 \\ 14 \\ 17 \\ 16 \\ 15 \\ 8 \\ 11 \\ 8 \end{bmatrix}. \tag{14}$$

In order to create the $N = 12$ by $J + 1 = 3 + 1 = 4\mathbf{X}$ matrix, we have selected the fourth experimental group to be the contrasting group. The first column of $\mathbf{X}$ codes for the intercept and is composed only of 1s. For the other columns of $\mathbf{X}$, the values for the observations of the contrasting group will all be equal to $-1$. The second column of $\mathbf{X}$ will use values of 1 for the observations of the first group, the third column of $\mathbf{X}$ will use values of 1 for the observations of the second group, and the fourth column of $\mathbf{X}$ will use values of 1 for the observations of the third group:

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 \end{bmatrix}. \tag{15}$$

With this effect coding scheme, we obtain the following **b** vector of regression coefficients:

$$\mathbf{b} = \begin{bmatrix} 15 \\ 3 \\ 2 \\ 1 \end{bmatrix}. \quad (16)$$

We can check that the intercept is indeed equal to the grand mean (i.e., 15) and that the slope corresponds to the difference between the corresponding groups and the grand mean. When using the MLR approach to the ANOVA, the predicted values correspond to the group means, and this is indeed the case here.

### Alternatives

The two main alternatives to effect coding are dummy coding and contrast coding. Dummy coding is quite similar to effect coding, the only difference being that the contrasting group is always coded with values of 0 instead of −1. With dummy coding, the intercept is equal to the mean of the contrasting group, and each slope is equal to the mean of the contrasting group minus the mean of the group under consideration. For contrast coding, a set of (generally orthogonal, but linear independent is sufficient) $J$ contrasts is chosen for the last $J$ columns of **X.** The values of the intercept and slopes will depend upon the specific set of contrasts used.

*Hervé Abdi*

***See also*** Analysis of Covariance (ANCOVA); Analysis of Variance (ANOVA); Contrast Analysis; Dummy Coding; Mean Comparisons; Multiple Regression

#### Further Readings

Abdi, H., Edelman, B., Valentin, D., & Dowling, W. J. (2009). *Experimental design and analysis for psychology.* Oxford, UK: Oxford University Press.

Edwards, A. L. (1985). *Data analysis for research designs.* New York: Freeman.

Edwards, A. L. (1985). *Multiple regression analysis and the analysis of variance and covariance.* New York: Freeman.

Fox, J. (2008). *Applied regression analysis and generalized linear models.* Thousand Oaks, CA: Sage.

## Effect Size, Measures of

*Effect size* is a statistical term for the measure of associations between two variables. It is widely used in many study designs, such as meta-analysis, regression, and analysis of variance (ANOVA). The presentations of effect size in these study designs are usually different. For example, in meta-analysis—an analysis method for combining and summarizing research results from different studies—effect size is often represented as the standardized difference between two continuous variables' means. In analysis of variance, effect size can be interpreted as the proportion of variance explained by a certain effect versus total variance. In each study design, due to the characteristic of variables, say, continuous versus categorical, there are several ways to measure the effect size. This entry discusses the measure of effect size by different study designs.

### Measures and Study Designs

#### *Meta-Analysis*

Meta-analysis is a study of methodology to summarize results across studies. Effect size was introduced as standardized mean differences for continuous outcome. This is especially important for studies that use different scales. For example, in a meta-analysis for the study of different effects for schizophrenia from a drug and a placebo, researchers usually use some standardized scales to measure patients' situation. These scales can be the Positive and Negative Syndrome Scale (PANSS) or the Brief Psychiatric Rating Scale (BPRS). The PANSS is a 30-item scale, and scores range from 30 to 210. The BPRS scale is a 16-item scale, and one can score from 16 to 112. Difference studies may report results measured on either scale. When a researcher needs to use meta-analysis to combine studies with both scales reported, it would be better for him or her to convert those study results into a common standardized score so that those study results become comparable. Cohen's $d$ and Hedge's $g$ are common effect sizes used in meta-analysis with continuous outcomes.

For a dichotomized outcome, the odds ratio is often used as an indicator of effect size. For

example, a researcher may want to find out whether smokers have greater chances of having lung cancer compared to nonsmokers. He or she may do a meta-analysis with studies reporting how many patients, among smokers and nonsmokers, were diagnosed with lung cancer. The odds ratio is appropriate to use when the report is for a single study. One can compare study results by investigating odds ratios for all of these studies.

The other commonly used effect size in meta-analysis is the correlation coefficient. It is a more direct approach to tell the association between two variables.

#### Cohen's *d*

Cohen's $d$ is defined as the population means difference divided by the common standard deviation. This definition is based on the $t$-test on means and can be interpreted as the standardized difference between two means. Cohen's $d$ assumes equal variance of the two populations. For two independent samples, it can be expressed as $d = \frac{m_A - m_B}{\sigma}$ for a one-tailed effect size index and $d = \frac{|m_A - m_B|}{\sigma}$ for a two-tailed effect size index. Here, $m_A$ and $m_B$ are two population means in their raw scales, and $\sigma$ is the standard deviation of either population (both population means have equal variance). Because the population means and standard deviations are usually unknown, sample means and standard deviations are used to estimate Cohen's $d$. One-tailed and two-tailed effect size index for $t$-test of means in standard units are $d = \frac{\overline{x}_A - \overline{x}_B}{s}$ and $d = \frac{|\overline{x}_A - \overline{x}_B|}{s}$, where $\overline{x}_A$ and $\overline{x}_B$ are sample means, and $s$ is the common standard deviation of both samples.

For example, a teacher wanted to know whether the ninth-grade boys or the ninth-grade girls in her school were better at reading and writing. She randomly selected 10 boys and 10 girls from all ninth-grade students and obtained the reading and writing exam score means of all boys and girls, say, 67 and 71, respectively. She found that the standard deviations of both groups were 15. Then for a two-tailed test, the effect size is

$$d = \frac{|\overline{x}_A - \overline{x}_B|}{s} = \frac{|67 - 71|}{15} = 0.27.$$

Cohen used the terms *small, medium,* and *large* to represent relative size of effect sizes. The corresponding numbers are 0.2, 0.5, and 0.8, respectively. In the example above, the effect size of 0.27 is small, which indicates that the difference between reading and writing exam scores for boys and girls is small.

#### Glass's *g*

Similarly, Glass proposed an effect size estimator using a control group's standard deviation to standardize mean difference: $g' = \frac{\overline{x}^E - \overline{x}^C}{s^C}$. Here, $\overline{x}^E$ and $\overline{x}^C$ are the sample means of an experimental group and a control group, respectively, and $s^C$ is the standard deviation of the control group.

This effect size assumes multiple treatment comparisons to the control group, and that treatment standard deviations differ from each other.

#### Hedge's *g*

However, neither Cohen's $d$ nor Glass's $g$ takes the sample size into account, and the equal population variance assumption may not hold. Hedge proposed a modification to estimate effect size as

$$g = \frac{\overline{x}^E - \overline{x}^C}{s}, \text{ where}$$
$$s = \sqrt{\frac{(n^E - 1)(s^E)^2 - (n^C - 1)(s^C)^2}{n^E + n^C - 2}}.$$

Here, $n^E$ and $n^C$ are sample sizes of treatment and control, and $s^E$ and $s^C$ are sample standard deviations of treatment and control. Comparing to above effect size estimators, Hedges's $g$ uses pooled sample standard deviations to standardize mean difference.

However, the above estimator has a small sample bias. An approximate unbiased estimator of effect size defined by Hedges and Olkin is

$$g = \frac{\overline{x}^E - \overline{x}^C}{s}\left(1 - \frac{3}{4N - 9}\right), \text{ where}$$
$$s = \sqrt{\frac{(n^E - 1)(s^E)^2 - (n^C - 1)(s^C)^2}{n^E + n^C - 2}}.$$

Here, $N$ is the total sample size of both groups. Especially when sample sizes in treatment and control are equal, this estimator is the unique minimum variance unbiased estimator. Many meta-analysis software packages like Metawin use

Hedge's $g$ as the default effect size for continuous outcomes.

#### Odds Ratio

Odds ratio is a commonly used effect size for categorical outcomes. Odds ratio is the ratio of odds in Category 1 versus odds in Category 2. For example, a researcher wanted to find out the relationship between smoking and getting lung cancer. He recruited two groups of subjects: smokers and nonsmokers. After a few years of following up, he found that there were $N_{11}$ subjects diagnosed with lung cancer among smokers and $N_{21}$ subjects among nonsmokers. There were $N_{12}$ and $N_{22}$ subjects who didn't have lung cancer. The odds of having lung cancer among smokers and nonsmokers are estimated as $\frac{N_{11}}{N_{12}}$ and $\frac{N_{21}}{N_{22}}$, respectively. The odds ratio of having lung cancer in smokers compared to nonsmokers is the ratio of the above two odds, which is $\frac{N_{11}}{N_{12}} \Big/ \frac{N_{21}}{N_{22}} = \frac{N_{11}N_{22}}{N_{12}N_{21}}$.

The scale is different from the effect sizes of continuous variables, such as Cohen's $d$ and Hedge's $g$, so it is not appropriate to compare the size of the odds ratio with the effect sizes described above.

#### Pearson Correlation Coefficient ($r$)

The Pearson correlation coefficient ($r$) is also a popular effect size. It was first introduced by Karl Pearson to measure the strength of the relationship between two variables. The range of the Pearson correlation coefficient is from $-1$ to 1. Cohen gave general guidelines for the relative sizes of the Pearson correlation coefficient as small, $r = 0.1$; medium, $r = 0.3$; and large, $r = 0.5$. Many statistical packages, such as SAS and IBM® SPSS® (PASW) 18.0 (an IBM company, formerly called PASW® Statistics), and Microsoft's Excel can compute the Pearson correlation coefficient.

Many meta-analysis software packages, such as Metawin and Comprehensive Meta-Analysis, allow users to use correlations as data and calculate effect size with Fisher's $z$ transformation. The transformation formula is $z = \frac{1}{2}\ln(\frac{1+r}{1-r})$, where $r$ is the correlation coefficient.

The square of the correlation coefficient is also an effect size used to measure how much variance is explained by one variable versus the total variance. It is discussed in detail below.

### *Analysis of Variance*

There are some other forms of effect size used to measure the magnitude of effects. These effect sizes are often used in analysis of variance (ANOVA). They measure how much variance is introduced by a new explanatory variable, and they are ratios of extra variance caused by an explanatory variable versus total variance. These measures include squared correlation coefficient and correlation ratio, eta-squared and partial eta-squared, omega squared, and intraclass correlation. The following paragraphs give a brief discussion of these effect sizes. For easier understanding, one-way ANOVA is used as an example.

#### Squared Correlation Coefficient ($r^2$)

In regression analysis, the squared correlation coefficient is a commonly used effect size based on variance. The expression is $r^2 = \frac{\sigma_T^2 - \sigma_{RL}^2}{\sigma_T^2}$, where $\sigma_T^2$ is the total variance of the dependent variable and $\sigma_{RL}^2$ is the variance explained by other variables. The range of the squared correlation coefficient is from 0 to 1. It can be interpreted as the proportion of variance shared by two variables. For example, an $r^2$ of 0.35 means that 35% of the total variance is shared by two variables.

#### Eta-Squared ($\eta^2$) and Partial Eta-Squared ($\eta_p^2$)

Eta-squared and partial eta-squared are effect sizes used in ANOVAs to measure degree of association in a sample. The effect can be the main effect or interaction in an analysis of variance model. It is defined as the sum of squares of the effect versus the sum of squares of the total. Eta-squared can be interpreted as the proportion of variability caused by that effect for the dependent variable. The range of an eta-squared is from 0 to 1. Suppose there is a study of the effects of education and experience on salary, and that the eta-squared of the education effect is 0.35. This means that 35% of the variability in salary was caused by education.

Eta-squared is additive, so the eta-squared of all effects in an ANOVA model sums to 1. All effects include all main effects and interaction effects, as well as the intercept and error effect in an ANOVA table.

Partial eta-squared is defined as the sum of squares of the effect versus the sum of squares of the effect plus error. For the same effect in the same study, partial eta-squared is always larger than eta-squared. This is because the denominator in partial eta-squared is smaller than that in eta-squared. For the previous example, the partial eta-squared may be 0.49 or 0.65; it cannot be less than 0.35. Unlike eta-squared, the sum of all effects' partial eta-squared may not be 1, and in fact can be larger than 1.

The statistical package SPSS will compute and print out partial eta-squared as the effect size for analysis of variance.

#### Omega Squared ($\omega^2$)

Omega squared is an effect size used to measure the degree of association in fixed and random effects analysis of variance study. It is the relative reduction variance caused by an effect. Unlike eta-squared and partial eta-squared, omega squared is an estimate of the degree of association in a population, instead of in a sample.

#### Intraclass Correlation ($\rho_I^2$)

Intraclass correlation is also an estimate of the degree of association in a population in random effects models, especially in psychological studies. The one-way intraclass correlation coefficient is defined as the proportion of variance of random effect versus the variance of this effect and error variance. One estimator is $\hat{\rho}_I^2 = \frac{MS_{\text{effect}} - MS_{\text{error}}}{MS_{\text{effect}} + df_{\text{effect}} MS_{\text{error}}}$, where $MS_{\text{effect}}$ and $MS_{\text{error}}$ are mean squares of the effect and error, that is, the mean squares of between-group and within-group effects.

### *Other Effect Sizes*

#### $R^2$ in Multiple Regression

As with the other effect sizes discussed in the regression or analysis of variance sections, $R^2$ is a statistic used to represent the portion of variance explained by explanatory variables versus the total variance. The range of $R^2$ is from 0, meaning no relation between the dependent variable and explanatory variables, to 1, meaning all variances can be explained by explanatory variables.

#### $\omega$ and Cramer's $V$

The effect sizes $\omega$ and Cramer's $V$ are often used for categorical data based on *chi-square*. Chi-square is a nonparametric statistic used to test potential difference among two or more categorical variables. Many statistical packages, such as SAS and SPSS, show this statistic in output.

The effect size $\omega$ can be calculated from chi-square and total sample size $N$ as $\overline{\omega} = \sqrt{\frac{\chi^2}{N}}$. However, this effect size is used only in the circumstances of $2 \times 2$ contingency tables. Cohen gave general guidelines for the relative size of $\omega$: 0.1, 0.3, and 0.5 represent small, medium, and large effect sizes, respectively.

For a table size greater than $2 \times 2$, one can use Cramer's $V$ (sometimes called Cramer's $\varphi$) as the effect size to measure the strength of association. Popular statistical software such as SAS and SPSS can compute this statistic. One can also calculate it from the chi-square statistic using the formula $V = \sqrt{\frac{\chi^2}{N*L}}$, where $N$ is the total sample size and $L$ equals the number of rows minus 1 or the number of columns minus 1, whichever is less. The effect size of Cramer's $V$ can be interpreted as the average multiple correlation between rows and columns. In a $2 \times 2$ table, Cramer's $V$ is equal to the correlation coefficient. Cohen's guideline for $\omega$ is also appropriate for Cramer's $V$ in $2 \times 2$ contingency tables.

## Reporting Effect Size

Reporting effect size in publications, along with the traditional null hypothesis test, is important. The null hypothesis test tells readers whether an effect exists, but it won't tell readers whether the results are replicable without reporting an effect size. Research organizations such as the American Psychological Association suggest reporting effect size in publications along with significance tests. A general rule for researchers is that they should at least report descriptive statistics such as mean and standard deviation. Thus, effect size can be calculated and used for meta-analysis to compare with other studies.

There are many forms of effect sizes. One must choose to calculate appropriate effect size based on the purpose of the study. For example, Cohen's *d* and Hedge's *g* are often used in meta-analysis to compare independent variables' means. In meta-analysis with binary outcomes, the odds ratio is often used to combine study results. The correlation coefficient is good for both continuous and categorical outcomes. To interpret variances explained by effect(s), one should choose effect sizes from $r^2$, eta-squared, omega squared, and so on.

Many popular statistical software packages can compute effect sizes. For example, meta-analysis software such as Metawin and Comprehensive Meta-Analysis can compute Hedge's *g*, odds ratio, and the correlation coefficient based on data type. Statistical software such as SPSS gives eta-squared as the effect size in analysis of variance procedures. Of course, one can use statistical packages such as SAS or Excel to calculate effect size manually based on available statistics.

*Qiaoyan Hu*

***See also*** Analysis of Variance (ANOVA); Chi-Square Test; Correlation; Hypothesis; Meta-Analysis

**Further Readings**

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlando, FL: Academic Press.

Iversen, G. R., & Norpoth, H. (1976). *Analysis of variance* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–001). Beverly Hills, CA: Sage.

Kirk, R. E. (1982). *Experimental design: Procedures for the behavioral sciences* (2nd ed.). Belmont, CA: Brooks/Cole.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* Thousand Oaks, CA: Sage.

Sutton, A. J., Abrams, K. R., Jones, D. R., Sheldon, T. A., & Song, F. (2000). *Methods for meta-analysis in medical research* (2nd ed.). London: Wiley.

Volker, M. A. (2006). Reporting effect size estimates in school psychology research. *Psychology in the Schools, 43*(6), 653–672.

# Endogenous Variables

Endogenous variables in causal statistical modeling are variables that are hypothesized to have one or more variables at least partially explaining them. Commonly referred to in econometrics and the structural equation modeling family of statistical techniques, endogenous variables may be effect variables that precede other endogenous variables; thus, although some consider endogenous variables to be dependent, such a definition is technically incorrect.

Theoretical considerations must be taken into account when determining whether a variable is endogenous. Endogeneity is a property of the model, not the variable, and will differ among models. For example, if one were to model the effect of income on adoption of environmental behaviors, the behaviors would be endogenous, and income would be exogenous. Another model may consider the effect of education on income; in this case, education would be exogenous and income endogenous.

## The Problem of Endogeneity

One of the most commonly used statistical models is ordinary least squares regression (OLS). A variety of assumptions must hold for OLS to be the best unbiased estimator, including the independence of errors. In regression models, problems with endogeneity may arise when an independent variable is correlated with the error term of an endogenous variable. When observational data are used, as is the case with many studies in the social sciences, problems with endogeneity are more prevalent. In cases where randomized, controlled experiments are possible, such problems are often avoided.

Several sources influence problems with endogeneity: when the true value or score of a variable is not actually observed (measurement error), when a variable that affects the dependent variable is not included in the regression, and when recursivity exists between the dependent and independent variables (i.e., there is a feedback loop between the dependent and independent variables). Each of these sources may occur alone or in conjunction with other sources.

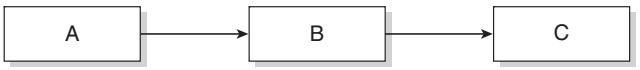



**Figure 1** Variables B and C Are Endogenous

The solution to problems with endogeneity is often to use instrumental variables. Instrumental variables methods include two-stage least squares, limited information maximum likelihood, and jackknife instrumental variable estimators. Advantages of instrumental variables estimation include the transparency of procedures and the ability to test the appropriateness of instruments and the degree of endogeneity. Instrumental variables are beneficial only when they are strongly correlated with the endogenous variable and when they are exogenous to the model.

### Endogenous Variables in Structural Equation Modeling

In structural equation modeling, including path analysis, factor analysis, and structural regression models, endogenous variables are said to be "downstream" of either exogenous variables or other endogenous variables. Thus, endogenous variables can be both cause and effect variables.

Consider the simple path model in Figure 1. Variable A is exogenous; it does not have any variables causally prior to it in the model. B is endogenous; it is affected by the exogenous variable A while affecting C. C is also an endogenous variable, directly affected by B and indirectly affected by A.

As error associated with the measurement of endogenous variables can bias standardized direct effects on endogenous variables, structural equation modeling uses multiple measures of latent constructs in order to address measurement error.

*Kristin Floress*

***See also*** Exogenous Variables; Latent Variable; Least Squares, Methods of; Regression to the Mean; Structural Equation Modeling

**Further Readings**

Bound, J., Jaeger, D. A., & Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the explanatory variable is weak. *Journal of the American Statistical Association, 90,* 443–450.

Kennedy, P. (2008). *A guide to econometrics.* Malden, MA: Blackwell.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford.

## Error

Error resides on the statistical side of the fault line separating the deductive tools of mathematics from the inductive tools of statistics. On the mathematics side of the chasm lays perfect information, and on the statistics side exists estimation in the face of uncertainty. For the purposes of estimation, error describes the unknown, provides a basis for comparison, and serves as a hypothesized placeholder enabling estimation. This entry discusses the role of error from a modeling perspective and in the context of regression, ordinary least squares estimation, systematic error, random error, error distributions, experimentation, measurement error, rounding error, sampling error, and nonsampling error.

### Modeling

For practical purposes, the universe is stochastic. For example, any "true" model involving gravity would require, at least, a parameter for every particle in the universe. One application of statistics is to quantify uncertainty. Stochastic or probabilistic models approximate relationships within some locality that contains uncertainty. That is, by holding some variables constant and constraining others, a model can express the major relationships of interest within that locality and amid an acceptable amount of uncertainty. For example, a model describing the orbit of a comet around the sun might contain parameters corresponding to the large bodies in the solar system and account for all remaining gravitational pulls with an error term.

Model equations employ error terms to represent uncertainty or the negligible contributions. Error terms are often additive or multiplicative placeholders, and models can have multiple error terms.

*Additive:* $E = MC^2 + \varepsilon$, where $\varepsilon$ is an error term perfecting the equation

*Multiplicative:* $y = \alpha + x^{\beta}\varepsilon$

*Other:* $y = e^{\beta(x+\varepsilon_{ME})} + \varepsilon$, where $\varepsilon_{ME}$ is measurement error corresponding to $x$, and $\varepsilon$ is an additive error term.

## Development of Regression

The traditional modeling problem is to solve a set of inconsistent equations—characterized by the presence of more equations than unknowns. Early researchers cut their teeth on estimating physical relationships in astronomy and geodesy—the study of the size and shape of the earth—expressed by a set of $k$ inconsistent linear equations of the following form:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} \\ y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} \\ &\vdots \\ y_k &= \beta_0 + \beta_1 x_{k1} + \beta_2 x_{k2} + \cdots + \beta_p x_{kp}, \end{aligned}$$

where the $x$s and $y$s are measured values and the $p + 1\beta$s are the unknowns.

Beginning in antiquity, these problems were solved by techniques that reduced the number of equations to match the number of unknowns. In 1750, Johann Tobias Mayer assembled his observations of the moon's librations into a set of inconsistent equations. He was able to solve for the unknowns by grouping equations and setting their sums equal to zero—an early step toward $\Sigma\varepsilon_i = 0$. In 1760, Roger Boscovich began solving inconsistent equations by minimizing the sum of the absolute errors ($\Sigma|\varepsilon_i| = 0$) subject to an adding-up constraint; by 1786, Pierre-Simon Laplace minimized the largest absolute error; and later, Adrien-Marie Legendre and Carl Friedrich Gauss began minimizing the sum of the squared error terms ($\Sigma\varepsilon_i^2 = 0$) or the least squares. Sir Francis Galton, Karl Pearson, and George Udny Yule took the remaining steps in building least squares regression, which combines two concepts involving errors: Estimate the coefficients by minimizing $\Sigma\varepsilon_i^2 = 0$, and assume that $\varepsilon_{ij} \overset{IID}{\sim} N(o, \sigma^2)$. This progression of statistical innovations has culminated in a family of regression techniques incorporating a variety of estimators and error assumptions.

Statistical errors are the placeholders representing that which remains unquantified or inconsistent in a hypothesized relationship. In assuming that these inconsistencies behave reasonably, researchers are able to find reasonable solutions.

## Ordinary Least Squares Estimation

Ordinary least squares (OLS) estimates are derived from fitting one equation to explain a set of inconsistent equations. There are basically six assumptions implicit in OLS estimation, all of which regard errors as follows:

1. Misspecification error is negligible—the functional form is reasonable and no significant $x$s are absent from the model.
2. Least squares, $\Sigma\varepsilon_i^2 = 0$, estimation is reasonable for this application.
3. Measurement error is negligible—$y$ and the $x$s are accurately measured.
4. Error terms are independent.
5. Error terms are identically distributed.
6. Error terms are approximately normal with a mean of zero and the same variance.

The strength of the solution is sensitive to the underlying assumptions. In practice, assumptions are never proven, only disproven or failed to be disproven. This asymmetrical information regarding the validity of the assumptions proves to be hazardous in practice, leading to biased estimates.

The statistical errors, $\varepsilon_i$, in most models are estimated by residuals,

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_p x_{ip}.$$

Statistical errors are unobservable, independent, and unconstrained, whereas residuals are observable estimates of the statistical errors, not independent, and usually constrained to sum to zero.

## Systematic Errors Versus Random Errors

Error can be categorized in two types: systematic error—also known as statistical bias, lack-of-fit error, or fixed error—and random error. Systematic error describes a separate error pattern that is recognizable as distinct from all other remaining random error. It can be attributed to some effect and might be controlled or modeled. Two typical sources of systematic error arise from model-fitting (misspecification) problems and measurement error problems. Both can create significant identifiable error patterns.

Random error consists of the remaining unexplained uncertainty that cannot be attributed to any particular important factor. Random error comprises the vestiges after removing large discernable patterns. Hence, random error may appear to be homogeneous, when it is actually a mix of faint, indistinct heterogeneity.

## Error Distributions

Stochastic models include underlying assumptions about their error terms. As violations of the assumptions become more extreme, the results become less reliable. Robustness describes a model's reliability in the face of departures from these underlying assumptions.

In practice, more thought should be given to robustness and the validity of the assumptions. In particular, the assumption that the errors are normally distributed is dramatically overused. This distribution provides a convenient, long-tailed, symmetrical shape, and it is ubiquitous as suggested by the central limit theorem—introduced by Laplace, circa 1810. The central limit theorem holds that distributions of means and sums of random variables converge toward approximate normality, regardless of the underlying distribution of the random variable. The distribution of the errors is often approximately normal because it is a function of the other distributions in the equation, which are often approximately normal.

Regardless of its convenience and ubiquity, the normality assumption merits testing. If this assumption is untenable, then there are other practical options:

1. Transform the response or regressors so that the distribution of errors is approximately normal.
2. Consider other attention-worthy distributions for the errors, such as the extreme value distribution, the Poisson, the gamma, the beta, and so on. Other distributions can provide reliable results at the expense of convenience and familiarity.
3. Contemplate a nonparametric solution. Although these techniques might not be "distribution free," they are less sensitive, or at least sensitive in different ways to the underlying distribution.
4. Try a resampling approach, such as bootstrapping, to create a surrogate underlying distribution.

## Experimental Error

The objective of design of experiments is to compare the effects of treatments on similar experimental units. The basis of comparison is relative to the unexplained behavior in the response, which might be referred to as unexplained error. If the treatment means differ greatly relative to the unexplained error, then the difference is assumed to be statistically significant. The purpose of the design is to estimate more accurately both the treatment means and the unexplained variation in the response, $y$. This can be achieved through refining the design structure and/or increasing the number of experimental units. Conceptually,

$$y = \text{Treatment Structure} + \text{Design Structure} + \text{Error Structure}.$$

To illustrate the role of error in a designed experiment, consider the one-way analysis of variance (ANOVA) corresponding to

$$y_{ij} = \mu_i + \varepsilon_{ij}, \text{ where } \varepsilon_{ij} \overset{IID}{\sim} N(o, \sigma_\varepsilon^2),$$

where $i = 1 \ldots p$ (number of treatments) and $j = 1 \ldots n_{\text{i}}$ (sample size of the $i$th treatment). The most notable hypotheses are

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_p$$

$$H_a : \mu_i \neq \mu_j \quad i \neq j.$$

The usual test statistic for this hypothesis is $F = \frac{MS_{\text{treatment}}}{MS_{\text{error}}}$, which is the ratio of two estimators

of $\sigma_\varepsilon^2$. The numerator is $MS_{\text{treatment}} = \frac{\sum n_i(\bar{y}_{i.}-\bar{y}_{..})^2}{p-1}$, which is unbiased only if $H_0$ is true. The denominator is $MS_{\text{error}} = \frac{\sum (y_{ij}-\bar{y}_{i.})^2}{N-p}$, which is unbiased regardless of whether $H_0$ is true. George W. Snedecor recognized the value of this ratio and named it the $F$ statistic in honor of Ronald A. Fisher, who was chiefly responsible for its derivation.

Difference is relative. The $F$ test illustrates how error serves as a basis for comparison. If the treatment means, $\frac{\sum n_i(\bar{y}_{i.}-\bar{y}_{..})^2}{p-1}$, vary relatively more than the observations within each treatment, $\frac{\sum (y_{ij}-\bar{y}_{i.})^2}{N-p}$, then the statistician should infer that $H_0$ is false. That is, if the discernible differences between the treatment means are unusually large relative to the unknown, $\sigma_\varepsilon^2$, then the differences are more likely to be genuine. ANOVA is an analysis of means based on analyzing variances of errors.

## Measurement Error

Measurement error is the difference between the "true" value and the measured value. This is sometimes called observational error. For many models, one implied assumption is that the inputs and the outputs are measured accurately enough for the application. This is often false, especially with continuous variables, which can only be as accurate as the measurement and data storage devices allow. In practice, measurement error is often unstable and difficult to estimate, requiring multiple measurements or independent knowledge.

There are two negative consequences due to measurement error in the regressor, $x$. First, if the measurement error variance is large relative to the variability in $x$, then the coefficients will be biased. In a simple regression model for example, measurement error in $x$ will cause $\hat{\beta}_0$ to converge to a slightly larger value than $\beta_0$ and $\hat{\beta}_1$ to be "attenuated" that is, the measurement error shrinks $\hat{\beta}_1$ so that it will underestimate $\beta_1$. Second, if the measurement error in $x$ is large relative to the variability of $y$, then this will increase the widths of confidence intervals. Both of these problems interfere with the two primary objectives of modeling: coefficient estimation and prediction.

The best solution for both objectives is to reduce the measurement error. This can be accomplished in three ways:

1. Improve the measurement device, possibly through calibration.
2. Improve the precision of the data storage device.
3. Replace $x$ with a more accurate measure of the same characteristic, $x_M$.

The next most promising solution is to estimate the measurement error and use it to "adjust" the parameter estimates and the confidence intervals. There are three approaches:

1. Collect repeated measures of $x$ on the same observations, thereby estimating the variance of the measurement error, $\sigma_{ME}^2$, and using it to adjust the regression coefficients and the confidence intervals.
2. Calibrate $x$ against a more accurate measure, $x_M$, which is unavailable for the broader application, thereby estimating the variance of the measurement error, $\sigma_{ME}^2$.
3. Build a measurement error model based on a validation data set containing $y$ and $x$ alongside the more accurate and broadly unavailable $x_M$. As long as the validation data set is representative of the target population, the relationships can be extrapolated.

For the prediction problem, there is a third solution for avoiding bias in the predictions, yet it does not repair the biased coefficients or the ample confidence intervals. The solution is to ensure that the measurement error present when the model was built is consistent as the model is applied.

## Rounding Error

Rounding error is often voluntary measurement error. The person or system causing the rounding is now a second stage in the measurement device. Occasionally, data storage devices lack the same precision as the measurement device, and this creates rounding error. More commonly, people or software collecting the information fail to retain the full precision of the data. After the data are collected, it is common to find unanticipated

applications—the serendipity of statistics, wanting more precision.

Large rounding error, $\varepsilon_R$, can add unwelcome complexity to the problem. Suppose that $x_2$ is measured with rounding error, $\varepsilon_R$, then a model involving $x_2$ might look like this:

$$y = \beta_0 + \beta_1 x_1 + \beta_2(x_2 + \varepsilon_R) + \varepsilon.$$

### Sampling and Nonsampling Error

The purpose of sampling is to estimate characteristics (mean, variance, etc.) of a population based upon a randomly selected representative subset. The difference between a sample's estimate and the population's value is due to two sources of error: sampling error and nonsampling error. Even with perfect execution, there is a limitation on the ability of the partial information contained in the sample to fully estimate population characteristics. This part of the estimation difference is due to sampling error—the minimum discrepancy due to observing a sample instead of the whole population. Nonsampling error explains all remaining sources of error, including nonresponse, selection bias, measurement error (inaccurate response), and so on, that are related to execution.

Sampling error is reduced by improving the sample design or increasing the sample size. Nonsampling error is decreased through better execution.

*Randy J. Bartlett*

***See also*** Error Rates; Margin of Error; Missing Data, Imputation of; Models; "Probable Error of a Mean, The"; Random Error; Residual Plot; Residuals; Root Mean Square Error; Sampling Error; Standard Deviation; Standard Error of Estimate; Standard Error of Measurement; Standard Error of the Mean; Sums of Squares; Systematic Error; Type I Error; Type II Error; Type III Error; Variability, Measure of; Variance; White Noise

#### Further Readings

Harrell, F. E., Jr. (2001). *Regression modeling strategies, with applications to linear models, logistic regression, and survival analysis*. New York: Springer.

Heyde, C. C., & Seneta, E. (2001). *Statisticians of the centuries*. New York: Springer-Verlag.

Kutner, M., Nachtsheim, C., Neter, J., & Li, W. (2005). *Applied linear statistical models* (5th ed.). Boston: McGraw-Hill Irwin.

Milliken, G., & Johnson, D. (1984). *Analysis of messy data: Vol. 1.*: Designed experiments. Belmont, CA: Wadsworth.

Snedecor, G. (1956). *Statistical methods applied to experiments in agriculture and biology* (5th ed.). Ames: Iowa State College Press.

Stigler, S. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Cambridge, MA: Belknap Press of Harvard University Press.

Weisberg, S. (1985). *Applied linear regression* (2nd ed.). New York: John Wiley.

## Error Rates

In research, *error rate* takes on different meanings in different contexts, including measurement and inferential statistical analysis. When measuring research participants' performance using a task with multiple trials, error rate is the proportion of responses that are incorrect. In this manner, error rate can serve as an important dependent variable. In inferential statistics, errors have to do with the probability of making a false inference about the population based on the sample data. Therefore, estimating and managing error rates are crucial to effective quantitative research.

This entry mainly discusses issues involving error rates in measurement. Error rates in statistical analysis are mentioned only briefly because they are covered in more detail under other entries.

### Error Rates in Measurement

In a task with objectively correct responses (e.g., a memory task involving recalling whether a stimulus had been presented previously), a participant's response can be one of three possibilities: no response, a correct response, or an incorrect response (error). Instances of errors across a series of trials are aggregated to yield error rate, ideally in proportional terms. Specifically, the number of errors divided by the number of trials in which one has an opportunity to make a correct response yields the error rate. Depending on the goals of the study, researchers

may wish to use for the denominator the total number of responses *or* the total number of trials (including nonresponses, if they are considered relevant). The resulting error rate can then be used to test hypotheses about knowledge or cognitive processes associated with the construct represented by the targets of response.

### *Signal Detection Theory*

One particularly powerful data-analytic approach employing error rates is Signal Detection Theory (SDT). SDT is applied in situations where the task involves judging whether a signal exists (e.g., "Was a word presented previously, or is it a new word?"). Using error rates in a series of trials in a task, SDT mathematically derives characteristics of participants' response patterns such as sensitivity (the perceived distinction between a signal and noise) and judgment criterion (the tendency to respond in one way rather than the other).

Typically, SDT is based on the following assumptions. First, in each trial, either a signal exists or it does not (e.g., a given word was presented previously or not). Even when there is no signal (i.e., the correct response would be negative), the *perceived* intensity of the stimuli varies randomly (caused by factors originating from the task or from the perceiver), which is called "noise." Noise follows a normal distribution with a mean of zero. Noise always accompanies signal, and because noise is added to signal, the distribution of *perceived* intensity of signal has the same (normal) shape. Each perceiver is assumed to have an internal set criterion (called threshold) used to make decisions in the task. If the perceived intensity (e.g., subjective familiarity) of the stimulus is stronger than the threshold, the perceiver will decide that there is a signal (respond affirmatively—e.g., indicate that the word was presented previously); otherwise, the perceiver will respond negatively. When the response is not consistent with the objective properties of the stimulus (e.g., a negative response to a word that *was* presented previously or an affirmative response to a word that was not), it is an error.

Responses are categorized into four groups: hits (signal is present and the perceiver responds affirmatively); misses (signal is present and the perceiver responds negatively); correct rejections (signal is not present and the perceiver responds negatively); and false alarms (signal is not present and the perceiver responds affirmatively). As indicated above, misses and false alarms are errors. Miss rate is calculated as the ratio of missed trials to the total number of trials with signal. False alarm rate is the ratio of trials with false alarms to the total number of trials without signal. Hit rate and miss rate sum to 1, as do correct rejection rate and false alarm rate.

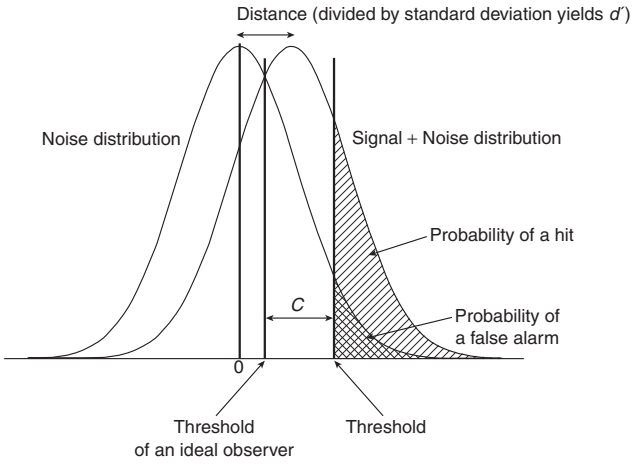



**Figure 1** Distributions of Signal and Noise

The objective of SDT is to estimate two indexes of participants' response tendencies from the error rates. The *sensitivity* (or discriminability) index ($d'$) pertains to the strength of the signal (or a perceiver's ability to discern signal from noise), and *response bias* (or strategy) ($C$) is the tendency to respond one way or the other (e.g., affirmatively rather than negatively). The value of $d'$ reflects the distance between the two distributions relative to their spread, so that a larger value means the signal is more easily discerned (e.g., in the case of word learning, this may imply that the learning task was effective). $C$ reflects the threshold of the perceiver minus that of an ideal observer. When the value of $C$ is positive, the perceiver is said to be *conservative* (i.e., requiring stronger intensity of the stimulus to respond affirmatively), and a perceiver with a negative $C$ is *liberal*. As the $C$ value increases, both miss rate and correct rejection rate increase (i.e., more likely to respond negatively both when there is a signal and when there is not); conversely, as it decreases, both hit rate and false

alarm rate increase. Bias is sometimes expressed as $\beta$, which is defined as the likelihood ratio of the signal distribution to noise distribution at the criterion (i.e., the ratio of the height of the signal curve to the height of the noise curve at the value of the threshold) and is equal to $e^{d' * ; C}$. The value would be greater than 1 when the perceiver is conservative and less than 1 when liberal. Sensitivity and bias can be estimated using a normal distribution function from hit rate and false alarm rate; because the two rates are independent of each other, it is necessary to obtain both of them from data.

### *Process Dissociation Procedure*

Process Dissociation Procedure (PDP) is a method that uses error rates to estimate the separate contributions of controlled (intentional) and automatic (unintentional) processes in responses. In tasks involving cognitive processes, participants will consciously (intentionally) strive to make correct responses. But at the same time, there may also be influences of automatic processes that are beyond conscious awareness or control. Using PDP, the researcher can estimate the independent influences of controlled and automatic processes from error rates.

The influences of controlled and automatic processes may work hand in hand or in opposite directions. For example, in a typical Stroop color naming task, participants are presented with color words (e.g., "red") and instructed to name the colors of the words' lettering, which are either consistent (e.g., red) or inconsistent (e.g., green) with the words. In certain trials, the response elicited by the automatic process is the correct response as defined in the task. For example, when the stimulus is the word "red" in red lettering, the automatic process (to read the word) will elicit the response "red," which is the same as the one dictated by the controlled process (to name the color of the lettering). Such trials are called *congruent* trials, because controlled and automatic processes elicit the same response. In other trials, the response elicited by the automatic process is not the response required in the task. In our example, when the word "green" is presented in red lettering, the automatic process will elicit the response "green." In this case, the directions of the influences that controlled and automatic processes have on the response are opposite to each other; these trials are called *incongruent* trials. The goal of PDP is to estimate the probabilities that controlled and automatic processes affect responses.

### *Other Issues With Error Rates in Measurement*

#### Speed-Accuracy Trade-Off

In tasks measuring facility of judgments (e.g., categorizing words or images), either error rate or response latency can be used as the basis of analysis. If the researcher wants to use error rates, it is desirable to have time pressure in the task in order to increase error rates, so that larger variability in error rate can be obtained. Without time pressure, in many tasks, participants will make mostly accurate responses, and it will be hard to discern meaningful variability in response facility.

#### Problems Involving High Error Rates

When something other than error rate is measured (e.g., response latency), error rate may be high for some or all participants. As a consequence, there may be too few valid responses to use in analysis. To address this issue, if only a few participants have error rates higher than a set criterion (ideally discerned by a discontinuity in the frequency distribution of error rates), the researcher may remove his or her data. However, if it is a more prevalent trend, the task may be too difficult (in which case, the task may have to be made easier) or inappropriate (in which case, the researcher should think of a better way to measure the construct).

#### Error Rates as a Source of Error Variance

In measures with no objectively correct or incorrect responses (e.g., Likert-scale ratings of attitudes or opinions), errors can be thought of as the magnitude of inaccuracy of the measurement. For example, when the wording of questionnaire items or the scale of a rating trial is confusing, or when some of the participants have response sets (e.g., a tendency to arbitrarily favor a particular response option), the responses may not accurately reflect what is meant to be measured. If error variance caused by peculiarities of the measurement or of some participants is considerable, the reliability

of the measurement is dubious. Therefore, the researcher should strive to minimize these kinds of errors, and check the response patterns within and across participants to see if there are any nontrivial, systematic trends not intended.

### Errors in Statistical Inference

In statistical inference, the concept of error rates is used in null hypothesis significance testing (NHST) to make judgments of how probable a result is in a given population. Proper NHST is designed to minimize the rates of two types of errors: Type II and particularly Type I errors.

A Type I error (false positive) occurs when a rejected null hypothesis is correct (i.e., an effect is inferred when, in fact, there is none). The probability of a Type I error is represented by the $p$ value, which is assessed relative to an a priori criterion, $\alpha$. The conventional criterion is $\alpha = .05$; that is, when the probability of a Type I error ($p$) is less than .05, the result is considered "statistically significant." Recently, there has been a growing tendency to report the exact value of $p$ rather than merely stating whether it is less than $\alpha$. Furthermore, researchers are increasingly reporting effect size estimates and confidence intervals so that there is less reliance on a somewhat arbitrary, dichotomous decision based on the .05 criterion.

A Type II error (false negative) occurs when a retained null hypothesis is incorrect (i.e., no effect is inferred when, in fact, there is one). An attempt to decrease the Type II error rate (by being more liberal in saying there is an effect) also increases the Type I error rate, so one has to compromise.

*Sang Hee Park and Jack Glaser*

***See also*** Error; False Positive; Nonsignificance; Significance, Statistical

#### Further Readings

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language, 30,* 513–541.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. New York: Oxford University Press.

Sanders, A. F. (1998). *Elements of human performance: Reaction processes and attention in human skill*. Mahwah, NJ: Lawrence Erlbaum.

Wickens, T. D. (2001). *Elementary signal detection theory*. New York: Oxford University Press.

## ESTIMATION

Estimation is the process of providing a numerical value for an unknown quantity based on information collected from a sample. If a single value is calculated for the unknown quantity, the process is called *point estimation.* If an interval is calculated that is likely, in some sense, to contain the quantity, then the procedure is called *interval estimation,* and the interval is referred to as a confidence interval. Estimation is thus the statistical term for an everyday activity: making an educated guess about a quantity that is unknown based on known information. The unknown quantities, which are called parameters, may be familiar population quantities such as the population mean $\mu$, population variance $\sigma^2$, and population proportion $\pi$. For instance, a researcher may be interested in the proportion of voters favoring a political party. That proportion is the unknown parameter, and its estimation may be based on a small random sample of individuals. In other situations, the parameters are part of more elaborate statistical models, such as the regression coefficients $\beta_0, \beta_1, \ldots, \beta_p$ in a linear regression model

$$Y = \beta_0 + \sum_{j=1}^{p} x_j \beta_j + \varepsilon,$$

which relates a response variable $Y$ to explanatory variables $x_1, x_2, \ldots, x_p$.

Point estimation is one of the most common forms of statistical inference. One measures a physical quantity in order to estimate its value, surveys are conducted to estimate unemployment rates, and clinical trials are carried out to estimate the cure rate (risk) of a new treatment. The unknown parameter in an investigation is denoted by $\theta$, assumed for simplicity to be

a scalar, but the results below extend to the case that $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$ with $k > 1$.

To estimate $\theta$, or, more generally, a real-valued function of $\theta$, $\tau(\theta)$, one calculates a corresponding function of the observations, a *statistic,* $\delta = \delta(X_1, X_2, \ldots, X_n)$. An *estimator* is any statistic $\delta$ defined over the sample space. Of course, it is hoped that $\delta$ will tend to be close, in some sense, to the unknown $\tau(\theta)$, but such a requirement is not part of the formal definition of an estimator. The value $\delta(x_1, x_2, \ldots, x_n)$ taken on by $\delta$ in a particular case is the *estimate* of $\tau(\theta)$, which will be our educated guess for the unknown value. In practice, the compact notation $\hat{\delta}$ is often used for both estimator and estimate.

The theory of point estimation can be divided into two parts. The first part is concerned with methods for finding estimators, and the second part is concerned with evaluating these estimators. Often, the methods of evaluating estimators will suggest new estimators. In many cases, there will be an obvious choice for an estimator of a particular parameter. For example, the sample mean is a natural candidate for estimating the population mean; the median is sometimes proposed as an alternative. In more complicated settings, however, a more systematic way of finding estimators is needed.

## Methods of Finding Estimators

The formulation of the estimation problem in a concrete situation requires specification of the probability model, *P,* that generates the data. The model $P$ is assumed to be known up to an unknown parameter $\theta$, and $P = P_\theta$ is written to express this dependence. The observations $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ are postulated to be the values taken on by the random observable $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ with distribution $P_\theta$. Frequently, it will be reasonable to assume that each of the $X_i$s has the same distribution, and that the variables $X_1, X_2, \ldots, X_n$ are independent. This situation is called the independent, identically distributed (i.i.d.) case in the literature and allows for a considerable simplification in our model.

There are several general-purpose techniques for deriving estimators, including methods based on moments, least-squares, maximum-likelihood, and Bayesian approaches. The method of moments is based on matching population and sample moments, and solving for the unknown parameters. Least-squares estimators are obtained, particularly in regression analysis, by minimizing a (possibly weighted) difference between the observed response and the value predicted by the model.

The method of maximum likelihood is the most popular technique for deriving estimators. Considered for fixed $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ as a function of $\theta$, the joint probability density (or probability) $p_\theta(\boldsymbol{x}) = p_\theta(x_1, \ldots, x_n)$ is called the *likelihood* of $\theta$, and the value $\hat{\theta} = \hat{\theta}(X)$ of $\theta$ that maximizes $p_\theta(\boldsymbol{x})$ constitutes the *maximum likelihood estimator* (MLE) of $\theta$. The MLE of a function $\tau(\theta)$ is defined to be $\tau(\hat{\theta})$.

In Bayesian analysis, a distribution $\pi(\theta)$, called a *prior distribution,* is introduced for the parameter $\theta$, which is now considered a random quantity. The prior is a subjective distribution, based on the experimenter's belief about $\theta$, prior to seeing the data. The joint probability density (or probability function) of $\boldsymbol{X}$ now represents the conditional distribution of $\boldsymbol{X}$ given $\theta$, and is written $p(\mathbf{x} \mid \theta)$. The conditional distribution of $\theta$ given the data $\mathbf{x}$ is called the posterior distribution of $\theta$, and by Bayes's theorem, it is given by

$$\pi(\theta \mid \boldsymbol{x}) = \pi(\theta) p(\boldsymbol{x} \mid \theta) / m(\boldsymbol{x}), \qquad (1)$$

where $m(\boldsymbol{x})$ is the marginal distribution of $\boldsymbol{X}$, that is, $m(\boldsymbol{x}) = \int \pi(\theta) p(\boldsymbol{x} \mid \theta)\, d\theta$. The posterior distribution which combines prior information and information in the data, is now used to make statements about $\theta$. For instance, the mean or median of the posterior distribution can be used as a point estimate of $\theta$. The resulting estimators are called *Bayes estimators.*

### *Example*

Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. Bernoulli random variables, which take the value 1 with probability $\theta$ and 0 with probability $1 - \theta$. A Bernoulli process results, for example, from conducting a survey to estimate the unemployment rate, $\theta$. In this context, the value 1 denotes the responder was unemployed. The first moment (mean) of the distribution is $\theta$ and the likelihood function is given by

$$p_\theta(x_1, \ldots, x_n) = \theta^y (1 - \theta)^{n-y}, \quad 0 \le \theta \le 1,$$

where $y = \sum x_i$. The method of moments and maximum-likelihood estimates of $\theta$ are both $\hat{\theta} = y/n$, that is, the intuitive frequency-based estimate for the probability of success given $y$ successes in $n$ trials. For a Bayesian analysis, if the prior distribution for the parameter $\theta$ is a Beta distribution, Beta$(\alpha, \beta)$,

$$\pi(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}, \quad \alpha, \beta > 0$$

the posterior distribution for $\theta$, from (1), is

$$\pi(\theta \,|\, \boldsymbol{x}) \propto \theta^{\alpha+y-1}(1-\theta)^{n+\beta-y-1}.$$

The posterior distribution is also a Beta distribution, $\theta \,|\, \boldsymbol{x} \sim$ Beta $(\alpha + y, n + \beta = y)$, and a Bayes estimate, based on, for example, the posterior mean, is $\hat{\theta} = (\alpha + y)/(n + \beta - y)$.

## Methods of Evaluating Estimators

For any given unknown parameter, there are, in general, many possible estimators, and methods to distinguish between good and poor estimators are needed. The general topic of evaluating statistical procedures is part of the branch of statistics known as *decision theory*. The error in using the observable $\hat{\theta} = \hat{\theta}(\boldsymbol{X})$ to estimate the unknown $\theta$ is $\hat{\varepsilon} = \hat{\theta} - \theta$. This error forms the basis for assessing the performance of an estimator. A commonly used finite-sample measure of performance is the mean squared error (MSE). The MSE of an estimator $\hat{\theta}$ of a parameter $\theta$ is the function of $\theta$ defined by $\mathrm{E}\big(\hat{\theta}(\boldsymbol{X}) - \theta\big)^2$ where $E(\,\cdot\,)$ denotes the expected value of the expression in brackets. The advantage of the MSE is that it can be decomposed into a *systematic error* represented by the square of the bias $\mathrm{B}\big(\hat{\theta};\, \theta\big) = \mathrm{E}\,[\hat{\theta}(\boldsymbol{X})] - \theta$ and the intrinsic variability represented by the variance $V\big(\hat{\theta};\, \theta\big) = \mathrm{var}\big(\hat{\theta}(\boldsymbol{X})\big)$. Thus,

$$E\big(\hat{\theta} - \theta\big)^2 = B^2\big(\hat{\theta};\, \theta\big) + V\big(\hat{\theta};\, \theta\big).$$

An estimator whose bias $B\big(\hat{\theta};\, \theta\big) = 0$ is called *unbiased* and satisfies $E[\hat{\theta}(\boldsymbol{X})] = \theta$ for all $\theta$, so that, on average, it will estimate the right value. For unbiased estimators, the MSE reduces to the variance of $\hat{\theta}$.

The property of *unbiasedness* is an attractive one, and much research has been devoted to the study of unbiased estimators. For a large class of problems, it turns out that among all unbiased estimators, there exists one that uniformly minimizes the variance for all values of the unknown parameter, and which is therefore *uniformly minimum variance unbiased* (UMVU). Furthermore, one can specify a lower bound on the variance of *any* unbiased estimator of $\theta$, which can sometimes be attained. The result is the following version of the *information inequality*

$$\mathrm{var}\big(\hat{\theta}(\boldsymbol{X})\big) \geq 1/I(\theta), \tag{2}$$

where

$$I(\theta) = E\left\{\left[\frac{\partial}{\partial\theta}\log p_\theta(\boldsymbol{X})\right]^2\right\} \tag{3}$$

is the *information* (or Fisher information) that $\boldsymbol{X}$ contains about $\theta$. The bound can be used to obtain the (absolute) *efficiency* of an unbiased estimator $\hat{\theta}$ of $\theta$. This is defined as

$$e(\hat{\theta}) = \frac{1/I(\theta)}{V(\hat{\theta};\, \theta)}.$$

By Equation 2, the efficiency is bounded above by unity; when $e(\hat{\theta}) = 1$, for all $\theta$, $\hat{\theta}$ is said to be *efficient*. Thus, an efficient estimator, if it exists, is the UMVU, but the UMVU is not necessarily efficient. In practice, there is no universal method for deriving UMVU estimators, but there are, instead, a variety of techniques that can *sometimes* be applied.

Interestingly, unbiasedness is not essential, and a restriction to the class of unbiased estimators may rule out some very good estimators, including maximum likelihood. It is sometimes the case, for example, that a trade-off occurs between variance and bias in such a way that a small increase in bias can be traded for a larger decrease in variance, resulting in an improvement in MSE. In addition, finding a best unbiased estimator is not straightforward. For instance, UMVU estimators, or even any unbiased estimator, may not exist for a given $\tau(\boldsymbol{\theta})$; or the bound in Equation 2 may not be attainable, and one then has to decide if one's candidate for

best unbiased estimator is, in fact, optimal. Therefore, there is scope to consider other criteria also, and possibilities include *equivariance, minimaxity,* and *robustness*.

In many cases in practice, estimation is performed using a set of independent, identically distributed observations. In such cases, it is of interest to determine the behavior of a given estimator as the number of observations increases to infinity (i.e., *asymptotically*). The advantage of asymptotic evaluations is that calculations simplify and it is also more clear how to measure estimator performance. Asymptotic properties concern a sequence of estimators indexed by $n$, $\hat{\theta}_n$, obtained by performing the same estimation procedure for each sample size. For example, $\overline{X}_1 = X_1$, $\overline{X}_2 = (X_1 + X_2)/2$, $\overline{X}_3 = (X_1 + X_2 + X_3)/3$, and so forth. A sequence of estimators $\hat{\theta}_n$ is said to be asymptotically optimal for $\theta$ if it exhibits the following characteristics:

*Consistency:* It converges in probability to the parameter it is estimating, i.e., $P(|\hat{\theta}_n - \theta| < \varepsilon) \to 1$ for every $\varepsilon > 0$.

*Asymptotic normality:* The distribution of $n^{1/2}(\hat{\theta}_n - \theta)$ tends to a normal distribution with mean zero and variance $1/I_1(\theta)$, where $I_1(\theta)$ is the Fisher information in a single observation, that is, Equation 3 with $\boldsymbol{X}$ replaced by $X_1$.

*Asymptotic efficiency:* No other asymptotically normal estimator has smaller variance than $\hat{\theta}_n$.

The small-sample and asymptotic results above generalize to vector-valued parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_k)$ and estimation of real- or vector-valued functions of $\boldsymbol{\theta}$, $\tau(\boldsymbol{\theta})$. One also can use the asymptotic variance as a means of comparing two asymptotically normal estimators through the idea of *asymptotic relative efficiency* (ARE). The ARE of one estimator compared with another is the reciprocal of the ratio of their asymptotic (generalized) variances. For example, the ARE of median to mean when the $X$s are normal is $2/\pi \approx 0.64$, suggesting a considerable efficiency loss in using the median at this case.

It can be shown under general regularity conditions that if $\hat{\boldsymbol{\theta}}$ is the MLE, then $\tau(\hat{\boldsymbol{\theta}})$ is an asymptotically optimal (most efficient) estimator of $\tau(\boldsymbol{\theta})$. There are other asymptotically optimal estimators, such as Bayes estimators. The method of moments estimator is not, in general, asymptotically optimal but has the virtue of being quite simple to use.

In most practical situations, it is possible to consider the use of several different estimators for the unknown parameters. It is generally good advice to use various alternative estimation methods in such situations, these methods hopefully resulting in similar parameter estimates. If a single estimate is needed, it is best to rely on a method that possesses good statistical properties, such as maximum likelihood.

*Panagiotis Besbeas*

***See also*** Accuracy in Parameter Estimation; Confidence Intervals; Inference: Deductive and Inductive; Least Squares, Methods of; Root Mean Square Error; Unbiased Estimator

**Further Readings**

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis* (2nd ed.). New York: Springer-Verlag.

Bickel, P. J., & Doksum, K. A. (2001). *Mathematical statistics: Basic ideas and selected topics, Vol. 1* (2nd ed.). Upper Saddle River, Englewood Cliffs, NJ: Prentice Hall.

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Pacific Grove, CA: Duxbury.

Johnson, N. L., & Kotz, S. (1994–2005). *Distributions in statistics* (5 vols.). New York: Wiley.

Lehmann, E.L., & Casella, G. (1998). *Theory of point estimation.* New York: Springer-Verlag.

## ETA-SQUARED

Eta-squared is commonly used in ANOVA and $t$ test designs as an index of the proportion of variance attributed to one or more effects. The statistic is useful in describing how variables are behaving within the researcher's sample. In addition, because eta-squared is a measure of effect size, researchers are able to compare effects of grouping variables or treatment conditions across related studies. Despite these advantages, researchers need

to be aware of eta-squared's limitations, which include an overestimation of population effects and its sensitivity to design features that influence its relevance and interpretability. Nonetheless, many social scientists advocate for the reporting of the eta-squared statistic, in addition to reporting statistical significance.

This entry focuses on defining, calculating, and interpreting eta-squared values, and will discuss the advantages and disadvantages of its use. The entry concludes with a discussion of the literature regarding the inclusion of eta-squared values as a measure of effect size in the reporting of statistical results.

## Defining Eta-Squared

Eta-squared ($\eta^2$) is a common measure of effect size used in *t* tests as well as univariate and multivariate analysis of variance (ANOVA and MANOVA, respectively). An eta-squared value reflects the strength or magnitude related to a main or interaction effect. Eta-squared quantifies the percentage of variance in the dependent variable (*Y*) that is explained by one or more independent variables (*X*). This effect tells the researcher what percentage of the variability in participants' individual differences on the dependent variable can be explained by the group or cell membership of the participants. This statistic is analogous to r-squared values in bivariate correlation ($r^2$) and regression analysis ($R^2$). Eta-squared is considered an additive measure of the unique variation in a dependent variable, such that nonerror variation is not accounted for by other factors in the analysis.

## Interpreting the Size of Effects

The value of $\eta^2$ is interpretable only if the *F* ratio for a particular effect is statistically significant. Without a significant *F* ratio, the eta-squared value is essentially zero and the effect does not account for any significant proportion of the total variance. Furthermore, some researchers have suggested cutoff values for interpreting eta-squared values in terms of the magnitude of the association between the independent and dependent measures. Generally, assuming a moderate sample size, eta-squared values of .09, .14, and .22 or greater could be described in the behavioral sciences as small, medium, and large. This index of the strength of association between variables has been referred to as *practical* significance. Determination of the size of effect based on an eta-squared value is largely a function of the variables under investigation. In behavioral science, *large effects* may be a relative term.

Partial eta-squared ($\eta_p^2$), a second estimate of effect size, is the ratio of variance due to an effect to the sum of the error variance and the effect variance. In a one-way ANOVA design that has just one factor, the eta-squared and partial eta-squared values are the same. Typically, partial eta-squared values are greater than eta-squared estimates, and this difference becomes more pronounced with the addition of independent factors to the design. Some critics have argued that researchers incorrectly use these statistics interchangeably. Generally, $\eta^2$ is preferred to $\eta_p^2$ for ease of interpretation.

## Calculating Eta-Squared

Statistical software programs, such as IBM® SPSS® (PASW) 18.0 (an IBM company, formerly called PASW® Statistics) and SAS, provide only the partial eta-squared values in the output, and not the eta-squared values. However, these programs provide the necessary values for the calculation of the eta-squared statistic. Using information provided in the ANOVA summary table in the output, eta-squared can be calculated as follows:

$$\eta^2 = \frac{SS_{\text{effect}}}{SS_{\text{total}}}.$$

When using between-subjects and within-subjects designs, the total sum of squares ($SS_{\text{total}}$) in the ratio represents the total variance. Likewise, the sum of squares of the effect (which can be a main effect or an interaction effect) represents the variance attributable to the effect. Eta-squared is the decimal value of the ratio and is interpreted as a percentage. For example, if $SS_{\text{total}} = 800$ and $SS_A = 160$ (the sum of squares of the main effect of *A*), the ratio would be .20. Therefore, the interpretation of the eta-squared value would be that the main effect of *A* explains 20% of the total variance of the dependent variable. Likewise, if $SS_{\text{total}} = 800$ and $SS_{A \times B} = 40$ (the sum of squares of the interaction between *A* and *B*), the ratio would be .05. Given these calculations, the

explanation would be that the interaction between $A$ and $B$ accounts for 5% of the total variance of the dependent variable.

### *Mixed Factorial Designs*

When using a mixed-design ANOVA, or a design that combines both between- and within-subject effects (e.g., pre- and posttest designs), researchers have differing opinions regarding whether the denominator should be the $SS_{\text{total}}$ when calculating the eta-squared statistic. An alternative option is to use the between-subjects variance ($SS_{\text{between subjects}}$) and within-subjects variance ($SS_{\text{within subjects}}$), separately, as the denominator to assess the strength of the between-subjects and within-subjects effects, respectively. Accordingly, when considering such effects separately, eta-squared values are calculated using the following formulas:

$$\eta^2 = \frac{SS_A}{SS_{\text{within subjects}}},$$

$$\eta^2 = \frac{SS_B}{SS_{\text{between subjects}}}, \quad \text{and}$$

$$\eta^2 = \frac{SS_{A \times B}}{SS_{\text{within subjects}}}.$$

When using $SS_{\text{between subjects}}$ and $SS_{\text{within subjects}}$ as separate denominators, calculated percentages are generally larger than when using $SS_{\text{total}}$ as the denominator in the ratio. Regardless of the approach used to calculate eta-squared, it is important to clearly interpret the eta-squared statistics for statistically significant between-subjects and within-subjects effects, respectively.

## Strengths and Weaknesses

### *Descriptive Measure of Association*

Eta-squared is a descriptive measure of the strength of association between independent and dependent variables in the sample. A benefit of the eta-squared statistic is that it permits researchers to descriptively understand how the variables in their sample are behaving. Specifically, the eta-squared statistic describes the amount of variation in the dependent variable that is shared with the grouping variable for a particular sample. Thus, because eta-squared is sample-specific, one disadvantage of eta-squared is that it may overestimate the strength of the effect in the population, especially when the sample size is small. To overcome this upwardly biased estimation, researchers often calculate an omega-squared ($\omega^2$) statistic, which produces a more conservative estimate. Omega-squared is an estimate of the dependent variable population variability accounted for by the independent variable.

### *Design Considerations*

In addition to the issue of positive bias in population effects, research design considerations may also pose a challenge to the use of the eta-squared statistic. In particular, studies that employ a multifactor completely randomized design should employ alternative statistics, such as partial eta- and omega-squared. In multifactor designs, partial eta-squared may be a preferable statistic when researchers are interested in comparing the strength of association between an independent and a dependent variable that excludes variance from other factors or when researchers want to compare the strength of association between the same independent and dependent measures across studies with distinct factorial designs. The strength of effects also can be influenced by the levels chosen for independent variables. For example, if researchers are interested in describing individual differences among participants but include only extreme groups, the strength of association is likely to be positively biased. Conversely, using a clinical research trial as an example, failure to include an untreated control group in the design might underestimate the eta-squared value. Finally, attention to distinctions between random and fixed effects, and the recognition of nested factors in multifactor ANOVA designs, is critical to the accurate use, interpretation, and reporting of statistics that measure the strength of association between independent and dependent variables.

## Reporting Effect Size and Statistical Significance

Social science research has been dominated by a reliance on significance testing, which is not

particularly robust to small ($N < 50$) or large sample sizes ($N > 400$). More recently, some journals publishers have adopted policies that require the reporting of effect sizes in addition to reporting statistical significance ($p$ values). In 2001, the American Psychological Association strongly encouraged researchers to include an index of effect size or strength of association between variables when reporting study results. Social scientists who advocate for the reporting of effect sizes argue that these statistics facilitate the evaluation of how a study's results fit into existing literature, in terms of how similar or dissimilar results are across related studies and whether certain design features or variables contribute to similarities or differences in effects. Effect size comparisons using eta-squared cannot be made across studies that differ in the populations they sampled (e.g., college students vs. elderly individuals) or in terms of controlling relevant characteristics of the experimental setting (e.g., time of day, temperature). Despite the encouragement to include strength of effects and significance testing, progress has been slow largely because effect size computations, until recently, were not readily available in statistical software packages.

*Kristen Fay and Michelle J. Boyd*

***See also*** Analysis of Variance (ANOVA); Effect Size, Measures of; Omega Squared; Partial Eta-Squared; $R^2$; Single-Subject Design; Within-Subjects Design

**Further Readings**

Cohen, J. (1973). Eta-squared and partial eta-squared in fixed factor ANOVA designs. *Educational and Psychological Measurement, 33,* 107–112.

Keppel, G. (1991). *Design and analysis: A researcher's handbook* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56,* 363–368.

Maxwell, S. E., & Delaney, H. D. (2000). *Designing experiments and analyzing data.* Mahwah, NJ: Lawrence Erlbaum.

Meyers, L. G., Gamst, G., & Guarino, A. J. (2006). *Applied multivariate research: Design and interpretation.* London: Sage.

Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology, 25,* 241–286.

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods, 8*(4), 434–447.

Pierce, C. A., Block, R. A., & Aguinis, H. (2004). Cautionary note on reporting eta-squared values from multifactor ANOVA designs. *Educational and Psychological Measurement, 64,* 916–924.

Thompson, B. (2002). "Statistical," "practical," and "clinical": How many kinds of significance do counselors need to consider? *Journal of Counseling and Development, 80*(1), 64–71.

# Ethics in the Research Process

In the human sciences, ethical concerns are felt at the level of the practicing scientist and are the focus of scholarly attention in the field of research ethics. Most of the ethical issues have to do with the scientist's obligations and the limits on permissible scientific activity. Perspectives on these issues are informed by ideas drawn from a variety of intellectual traditions, including philosophical, legal, and religious. Political views and cultural values also influence the interpretation of researcher conduct. Ethical questions about scientific activity were once considered external to the research endeavor, but today, it is taken for granted that researchers will reflect on the decisions that they make when designing a study and the ethical ramifications that their work might have. Scientists are also expected to engage in dialogue on topics that range from the controversial, such as the choice to study intelligence or conduct HIV trials, to the procedural, such as whether research volunteers are entitled to payment for their services.

## Key Themes in Research Ethics

Nearly any decision that scientists make can have ethical implications, but the questions most often addressed under the broad heading of Research Ethics can be grouped as follows: (a) guidelines and oversight, (b) autonomy and informed

consent, (c) standards and relativism, (d) conflicts of interest, and (e) the art of ethical judgment.

There is no exhaustive list of ethical problems because what constitutes an ethical problem for researchers is determined by a number of factors, including current fashions in research (not always in the human sciences) and such things as the prevailing political climate. Hence, a decision that a researcher makes might be regarded as controversial for several reasons, including a general sense that the action is out of step with the greater good. It is also common for decisions to be controversial because they are deemed to be contrary to the values that a particular scientific association promotes.

Legislation often piggybacks on such sentiments, with a close connection between views on what is ethical and what should (or should not) be enforced by law. In many countries, governmental panels weigh in on issues in research ethics. There, too, however, the categories of inquiry are fluid, with the panelists drawing on social, economic, and other considerations. The amorphous nature of the ethical deliberation that researchers might be party to thus results from there being so few absolutes in science or ethics. Most of the questions that researchers confront about the design of a study or proper conduct can readily be set against reasonable counter-questions. This does not rule out ethical distinctions, however. Just as the evaluation of scientific findings requires a combination of interpretive finesse and seasoned reflection, moral judgment requires the ability to critically evaluate supporting arguments.

## Guidelines and Oversight

Current codes of ethics have their origin in the aftermath of World War II, when interest in formal guidelines and oversight bodies first arose. News of the atrocities committed by Nazi researchers highlighted the fact that, with no consistent standards for scientific conduct, judgments about methods or approach were left to each researcher's discretion. The unscrupulous scientist was free to conduct a study simply to see what might happen, for example, or to conscript research "volunteers." The authors of the Nuremberg Code and, later, the Helsinki Declaration changed this by constructing a zone of protection out of several long-standing principles in law and medicine. In particular, from medical practice came injunctions like the doctor's "Do no harm." And from jurisprudence came the notion that nonconsensual touching can amount to assault, and that those so treated might have valid claims for restitution.

A common theme in these early codes was that research prerogatives must be secondary to the dignity and overall welfare of the humans under study. In the late 1970s, the *Belmont Report* expanded on this with its recommendation that a system of institutional review boards (IRBs) should ensure compliance with standards. The IRBs were also to be a liaison between researchers and anyone recruited by them. In addition, *Belmont* popularized a framework that researchers and scholars could use when discussing ethical issues. At first, the framework comprised moral principles like beneficence, justice, nonmaleficence, and respect for autonomy. In recent years, commentators have supplemented it with ideas from care ethics, casuistry, political theory, and other schools of thought.

Now commonplace in the lives of scientists, codes of ethics were once dismissed as mere attempts to "legislate morality." Skeptics also warned that scholarly debate about ethics would have little to offer researchers. In retrospect, it is plain that any lines between law and morality were blurred long before there was serious interest in codes of ethics. Not only that, rules and abstractions pertaining to ethics have always had to meet basic tests of relevance and practicality. In this process, scientists are not left out of the deliberations; they play an active role in helping to scrutinize the codes. Today, those codes are malleable artifacts, registers of current opinions about ethical values in science. They are also widely available online, usually with accompanying discussions of related ethical issues.

## Autonomy and Informed Consent

In human research, one group, made up of scientists, if not entire disciplines, singles out another group for study. Because this selection is almost never random, those in the latter group will usually know much less about the research and why they were selected. This can create a significant imbalance of power that can place the subjects

(also called recruits, patients, participants, or informants) in a subordinate position. The prospect that this will also leave the subjects vulnerable raises a number of ethical and legal issues to which researchers must respond when designing their studies.

Informed consent rules are the most common response to this problem of vulnerability. The specific details vary, but these rules usually call on researchers to provide a clear account of what might be in store for anyone who might serve as a subject. In most cases, the rules have a strong legalistic strain, in that the subjects *contract* to serve without pressure from the researchers, and with the option of changing their minds later. Such stipulations have been central to ethics codes from the outset, and in the review of research protocols.

Still, as informed consent rules are applied, their ethical significance rests with how researchers choose to define this operative concept. There are questions about how much or how little someone is consenting to when agreeing to participate, for example. There are questions about whether consent also gives the researcher license to disseminate study results in a particular manner. As might be expected, there are also differences of opinions on what it means to "inform" the subjects, and whether someone who is deliberately mis- or uninformed is thereby denied autonomy. What motivates these disagreements are the legitimate concerns about the degree of protection that is needed in some research, and the practicality of consent guidelines.

Whereas some scholars would have researchers err on the side of caution with very strict consent guidelines, others maintain that consent requirements can make certain types of research impossible or much less naturalistic. They argue that conditions worthy of study can be fleeting or sensitive enough that researchers have little time to negotiate the subjects' participation. Observational research into crowd behavior or the elite in government would be examples of studies that might be compromised if researchers were required to distribute informed consent paperwork first. In clinical trials, scientists might wonder how they should understand consent when their patients may be unable to truly understand the research.

There is much interest in developing research designs that can help the subjects feel more like collaborators and less like "guinea pigs," even without the ritualistic process of seeking consent. But commentators tend to agree that if consent rules are softened, there should be a presumption of unusually important benefits and minimal risks. As that demand is hard to meet, the trend is toward involving only bona fide volunteers. This preference for rather restrictive guidelines reflects apprehension about the legal consequences of involving people in research against their will, as well as a desire to avoid situations where a lack of consent might be a prelude to serious harm.

## Standards and Relativism

One of the oldest philosophical questions asks whether there is one moral truth or many truths. This question is of particular concern for researchers when there is what seems to be a "gray area" in ethical guidelines, that is, when a code of ethics does not appear to offer clear, explicit recommendations. In those settings, researchers ordinarily must decide which course represents the best compromise between their objectives and the interests of their subjects. Some commentators see the problem of *ethical relativism* in dilemmas like these. And although not everyone accepts the label of "relativism," with some preferring to speak of moral pluralism or even "flexibility" instead, most agree that the underlying problem bears on decisions about the design and conduct of research.

Research has never been governed by one overarching ethical standard that is complete in itself. Ethical standards have usually accommodated differences in methodological orientation, perceived levels of risk, and other criteria. Unfortunately, such accommodation has never been straightforward, and it is even less so now that disciplinary boundaries are shifting and traditional methods are being applied in novel ways. Psychologists now delve into what used to be considered medical research, historians work alongside public health specialists, investigative journalism bears similarities to undercover fieldwork, and some clinical studies rely on the analysis of patient narratives. Because of this, to recommend that anthropologists merely refer to the code of ethics for their discipline is to offer advice of very limited value.

Even if the disciplines were to hold static, the problem of relativism could surface wherever research takes place among populations whose beliefs about ethics or science differ from the researchers'. The researchers would still face the same questions about how they should adapt their own moral standards to those of their subjects, or vice versa. This problem is also complicated by the need for researchers to secure institutional approval for their studies. Because many of the points of reference in such oversight are subject to interpretation, a protocol might be rejected at one institution only to be accepted at another. Researchers are usually discouraged against "shopping around" for a more lenient audience. But it is an open question in many cases whether one review board has a better grasp of moral values than another.

Because it is, in some respects, the simplest alternative, there is some concern that laws and the fear of litigation will "solve" the relativism problem by ruling out even minor deviations from formal codes. On one hand, this could impose more standardization on the review of protocols. On the other hand, it could sharply limit professional discretion. For now, it is unclear how receptive the legal community is toward such efforts. It is also mistaken to presume that this would represent an unwarranted interference in scientific inquiry. Historically, researchers have not been left out of attempts to refine *legal* regulations, and the collaboration has helped to remove the ambiguity from codes that gave rise to some of the need to "relativize" ethical standards in the first place.

Concerns about relativism need not stall scientific activity or encourage researchers to simply devise rules of their own making. There are moral principles that transcend variations in the research setting. Experience suggests that prohibitions against causing physical harm to research subjects, for instance, or the falsifying of observational data, derive from such principles. Although these and other "bedrock" ethical principles are not enough to resolve all of the issues surrounding relativism, they can provide researchers with a baseline when adaptation of ethical guidelines seems necessary. Discussions about this adaptation in the scholarly literature, especially when informed by actual experience, also provide a resource, as do texts on research design.

## Conflicts of Interests

In research projects that stretch across several institutions, there might be disagreements over the allocation of credit for discoveries or published work. A researcher might also learn that the pharmaceutical corporation that is offering to fund his study would like some control over the reporting of the results. Another researcher might discover that her data stand to aid some people, even as they might threaten social programs on which others rely. Scientific organizations might express qualms about a study that could, despite the researcher's wishes, be used by governments to improve the interrogation of dissidents. Such conflicts of interest can reduce the incentive that researchers have to conduct their work, place subjects at risk, and undermine scientific credibility.

There is broad consensus that, in order to avoid this, the law should intervene to prevent conflicts that could lead to harm to subjects or society. Granting agencies are also expected to limit support for research that might create a conflict of interest or even give the appearance of one. Journal editors serve as gatekeepers, and they normally require that authors state compliance with ethics codes and declare any conflicts of interest. The remaining responsibility is left to researchers, educators, and others in supervisory roles. Through their contributions to the scholarly literature, researchers can help each other explore the issues surrounding conflicts of interest, and it is customary for training programs to incorporate these discussions.

Where these efforts lead to changes in codes of ethics, the goal is usually not to prevent them, but to provide researchers with a first line of defense. This still leaves researchers to rely very much on the individual ability to decide with which alliances or influences they are comfortable, and which ones will violate ethical principles. That may seem to offer little in the way of progress, but the fact remains that conflicts of interest are among the most complex ethical issues that researchers face, and there is no simple way to respond to them that will not create additional problems for the design of research or scientific innovation. In that respect, the researcher must perform a type of moral triage, where certain conflicts of interest are quickly dealt with or avoided,

and others are treated on an ongoing basis as part of the research process itself.

It is helpful to remember that in the days of the "Gentleman Scientist," there was little need to worry about conflicts of interest. Research was ethical if it conformed to a tacitly understood model of the humanistic intellectual. A similar model of virtue is still needed, yet today, researchers face conflicts of interest not covered by any code, written or otherwise, and they engage in science for any number of reasons. There are also good reasons to think that conflicts of interest can serve as test cases for the values that researchers are expected to support. Under that interpretation, an important benefit of the researcher's having to grapple with conflicts of interest would be the continual reexamination of such things as the role of business or government in science, or the value of knowledge for its own sake, aside from its practical applications.

## The Art of Ethical Judgment

Often, the researcher's most immediate ethical concern is the need to comply with institutional standards. This compliance is usually obtained by submitting a protocol in accordance with the various guidelines applicable to the type of research involved. This process can lend an administrative, official stamp to ethical assessment, but it can also obscure the wide range of values that are in play. In particular, critics charge that researchers too often feel pressured to look upon this assessment as something that can be reduced to a ledger of anticipated risks and benefits.

Critics also object that the review process rarely includes a provision for checking whether the risk-benefit forecast proves too accurate. Even if researchers did express interest in such verification, few mechanisms would enable it. The use of animals in research, as in a toxicity study, is said to illustrate some of these problems. Although risks and benefits are clearly at issue, training programs usually provide little advice on how researchers are to compare the risks to the animals against the benefits that patients are thought to gain from it. Concerns like these are of first importance, as scientists are taught early in their careers that protocols must be presented with assurances that a study is safe and in the public's interest.

It is also very common to criticize a clinical trial by alleging that the patients might not be fully apprised of the risks. Researchers who would design studies of genetic manipulation are asked to consider the damage that might result. And a prominent objection against including too many identifying details in a published ethnography is that the researcher will be unable to control the harm that this could cause. Where there are doubts about this language and the assessment that would declare one study ethical and another forbidden, there are questions about how well subjects are being protected or how research is being designed.

Most scholars would grant that researchers must do more than offer a cost-benefit analysis for their protocols. But this concession can still leave researchers without a clear sense of what that something else should be. For instance, textbooks on research design commonly recommend that researchers be honest and forthcoming with their subjects, aside from how this might affect the breakdown of anticipated risks. In practice, however, researchers ordinarily do not deal with only one or two moral principles. And even if this were not so, it can be unreasonable to ask that researchers give an account of how the risks from a loss of privacy in their subject population can be weighed against the benefits that a marginalized population might gain from a particular study.

In short, what is needed is an ability to identify variables that are either ignored or overemphasized in current assessment strategies. It is natural to turn to researchers to help refine that search. This is not asking that researchers develop moral wisdom. Researchers are, rather, the most qualified to devise improved methods of ethical assessment. They are also positioned best to bring any current deficiencies in those methods to the attention of fellow scholars. Needless to say, researchers have very practical reasons to improve their ability to justify their work: Society is unlikely to stop asking for an accounting of benefits and risks. And the emphasis on outcomes, whether risks, benefits, or some other set of parameters, is consistent with the priority usually given to empiricism in science. In other words, where there is even the possibility that researchers are unable to adequately gauge the effects of their work, there will be an impression that scientists are accepting significant shortcomings in the way that a study is deemed

a success or failure. That perception scientists cannot afford, so it will not do to fall back on the position that ambiguity about ethical values is simply the price that researchers must pay. More sensible is to enlist researchers in the search for ways to understand how the design of a study affects humans, animals, and the environment.

### Study Design

Initially considered something external to science, attention to ethical issues is now a major part of the design of studies, and scientific conduct in general. Although there is a popular misconception that this attention to ethical matters is unduly restrictive, that it places too many limits on the researcher's behavior, there is perhaps just as much effort toward elaborating on the obligations that the researcher has to study certain topics and provide much needed information about health and human behavior. The constraints that are placed on research design involve safeguards against unnecessarily manipulating, misleading, or otherwise harming the human participants. Researchers are also expected to envision the results that their studies might have, and to consider how what might seem minor decisions about methods or approach can have lasting effects on the general state of knowledge, as well as the perception that science has among the public. As a result, the decisions made when designing research will inevitably take place at the intersection of science and the changing attitudes about what is appropriate or necessary.

*Chris Herrera*

***See also*** Clinical Trial; Ethnography; Experimental Design; Informed Consent; Naturalistic Inquiry; Research Design Principles

#### Further Readings

Emanuel, E. J., Crouch, R. A., Arras, J. D., Moreno, J. D., & Grady, C. (Eds.). (2003). *Ethical and regulatory aspects of clinical research*. Baltimore: Johns Hopkins University Press.

Homan, R. (1991). *The ethics of social research*. New York: Longman.

Murphy, T. F. (2004). *Case studies in biomedical research ethics*. Cambridge: MIT Press.

Plomer, A. (2005). *The law and ethics of medical research: International bioethics and human rights*. London: Routledge.

van den Hoonaard, W. C. (Ed.). (2002). *Walking the tightrope: Ethical issues for qualitative researchers*. Toronto, Ontario, Canada: University of Toronto Press.

## ETHNOGRAPHY

Ethnography, in the simplest sense, refers to the writing or making of an abstract picture of a group of people. "Ethno" refers to people, and "graph" to a picture. The term was traditionally used to denote the composite findings of social science field-based research. That is, an ethnography represented a monograph (i.e., a written account) of fieldwork (i.e., the first-hand exploration of a cultural or social setting). In contemporary research, the term is used to connote the process of conducting fieldwork, as in "doing ethnography." For this entry, ethnography is addressed in the dual sense of monograph and research process.

### Traditions

Ethnography has been an integral part of the social sciences from the turn of the 20th century. The challenge in imparting an understanding of ethnography lies in not giving the impression there was or is a monolithic ethnographic way. A chronological, linear overview of ethnographic research would underrepresent the complexities and tensions of the historical development of ethnography. The development of ethnographic research cannot be neatly presented in periods or typologies, nor can ethnography be equated with only one academic discipline. Ethnography largely originated in the disciplines of anthropology and sociology; both anthropologists and sociologists have consistently based their research on intensive and extensive fieldwork. Ethnography, however, has evolved into different intellectual traditions for anthropologists and sociologists. By separately examining the disciplines of anthropology and sociology, one can gain an understanding of the origin of ethnography and how these disciplines have uniquely contributed to the foundation of ethnographic research.

### *Anthropology*

A primary concern of the discipline of anthropology is the study of culture, where culture is defined as the acquired meanings persons use to interpret experience and guide social behavior. In anthropology, an ethnography is a complex descriptive interpretation of a culture. Anthropologists historically ventured to remote, exotic settings to live among a people for a year or so to gain a firsthand understanding of their culture. Now, a dramatic difference between the ethnographer and "the other" is no longer a criterion of anthropological ethnography. The study of tribal or primitive cultures has evolved into the study of a wide range of cultural concerns, such as cultural events or scenes. Employing a cross-cultural perspective persists because it affords the ethnographer an ability to recognize aspects of human behavior capable of being observed, which is more likely to occur in the presence of differences than similarities. In anthropology, ethnographic fieldwork aims to discern cultural patterns of socially shared behavior. Anthropological ethnographers do not set out simply to observe culture; rather, they make sense of what they observe by making culture explicit. Conducting prolonged fieldwork, be it in a distant setting or a diverse one, continues to be a distinguishing characteristic of ethnography originating in anthropology.

There have been key differences, however, between American and British anthropologists' approach to ethnography. Generally, in the United States, ethnography and social anthropology were not separate anthropological pursuits. American anthropologists' pursuit of ethnographies of exotic groups and universal meanings governing human behavior were accommodated under the rubric "cultural anthropology." Conversely, British anthropologists historically have drawn a distinction between social anthropology and cultural anthropology. They distinguished between the ethnographer finely examining a specific group of people and the social anthropologist examining the same group to discern broad cultural patterns. Where these two disciplines intersect is that both traditions of anthropology have adhered to the standard of participant observation, both have been attentive to learning the native language, and both have been attuned to the constitution of everyday life.

### *Sociology*

Because of its characteristic firsthand exploration of social settings, ethnography is also deeply rooted in some intellectual traditions of the discipline of sociology. In sociology, ethnography entails studying social contexts and contexts of social action. The ethnographic study of small-scale urban and rural social settings dates back to the beginning of the twentieth century. Early ethnographies originating in sociology were unique for adhering to the scientific model of observation and data collection, for including quantitative techniques such as mapping, and for the use of literary devices of modern fiction, particularly in the United States. Moreover, sociological ethnographies helped define community as a product of human meaning and interaction.

Many early sociological ethnographies originated, in some manner, at the University of Chicago. During this generative era, the University of Chicago was considered to be on the forefront of sociology. More than half of the sociologists in the world were trained there, and a subgroup of these scholars created the Chicago School of ethnography. This group of sociologists proved to be prolific ethnographers and fundamentally shaped the discipline's embracing of ethnography. Chicago School ethnographies are marked by descriptive narratives portraying the face-to-face, everyday life in a modern, typically urban, setting. Central in these was the dynamic process of social change, such as rapid changes in values and attitudes.

### *Contemporary*

A current guiding assumption about ethnography is that ethnographic research can be conducted across place, people, and process as long as patterns of human social behavior are central. Both the methods and the resulting monograph are proving amenable across disciplines (e.g., economics, public policy), which is illustrated in the multiple theoretical perspectives and genres that are now represented in contemporary ethnography. Firsthand exploration offers a method of scientific inquiry attentive to diverse social contexts, and the

spectrum of contemporary theoretical frameworks affords a broad range of perspectives such as semiotics, poststructuralism, deconstructionist hermeneutics, postmodern, and feminist. Likewise, a range of genres of contemporary ethnographies has evolved in contrast to traditional ethnographies: autoethnography, critical ethnography, ethnodrama, ethnopoetics, and ethnofiction. Just as ethnography has evolved, the question of whether ethnography is doable across disciplines has evolved into how ethnographic methods might enhance understanding of the discipline-specific research problem being examined.

## Fieldwork

An expectation of ethnography is that the ethnographer goes into the field to collect his or her own data rather than rely on data collected by others. To conduct ethnography is to do fieldwork. Throughout the evolution of ethnography, fieldwork persists as the sine qua non. Fieldwork provides the ethnographer with a firsthand cultural/social experience that cannot be gained otherwise. Cultural/social immersion is irreplaceable for providing a way of seeing. In the repetitive act of immersing and removing oneself from a setting, the ethnographer can move between making up-close observations and then taking a distant contemplative perspective in a deliberate effort to understand the culture or social setting intellectually. Fieldwork provides a mechanism for learning the meanings that members are using to organize their behavior and interpret their experience.

Across discipline approaches, three fieldwork methods define and distinguish ethnographic research. Participant observation, representative interviewing, and archival strategies, or what Harry Wolcott calls experiencing, enquiring, and examining, are hallmark ethnographic methods. Therein, ethnographic research is renowned for the triangulation of methods, for engaging multiple ways of knowing. Ethnography is not a reductionist method, focused on reducing data into a few significant findings. Rather, ethnography employs multiple methods to flesh out the complexities of a setting. The specific methods used are determined by the research questions and the setting being explored. In this vein, Charles Frake advanced that the ethnographer seeks to find not only answers to the questions he or she brings into the field, but also questions to explain what is being observed.

### *Participant Observation*

Participant observation is the bedrock of doing and writing ethnography. It exists on a continuum from solely observing to fully participating. The ethnographer's point of entry and how he or she moves along the continuum is determined by the problem being explored, the situation, and his or her personality and research style. The ethnographer becomes "self-as-instrument," being conscious of how his or her level of participant observation to collect varying levels of data affects objectivity. Data are collected in fieldnotes written in the moment as well as in reflection later. Therein, participant observation can be viewed as a lens, a way of seeing. Observations are made and theoretically intellectualized at the limitation of not making other observations or intellectualizing observations with another theoretical perspective. The strength of extensive participant observation is that everyone, members and ethnographer, likely assumes natural behaviors over prolonged fieldwork. Repeated observations with varying levels of participation provide the means for how "ethnography makes the exotic familiar and the familiar exotic."

### *Interviewing*

A key assumption of traditional ethnographic research (i.e., anthropological, sociological) is that the cultural or social setting represents the sample. The members of a particular setting are sampled as part of the setting. Members' narratives are informally and formally elicited in relation to participant observation. Insightful fieldwork depends on both thoughtful, in-the-moment conversation and structured interviewing with predetermined questions or probes, because interviews can flesh out socially acquired messages. Yet interviewing is contingent on the cultural and social ethos, so it is not a given of fieldwork. Hence, a critical skill the ethnographer must learn is to discern when interviewing adds depth and understanding to participant observation and when interviewing interrupts focused fieldwork.

### *Archives*

Across research designs, using data triangulation (i.e., multiple data collection methods) can increase the scientific rigor of a study. Triangulation teases out where data converge and diverge. In ethnographic research where the ethnographer operates as the instrument in fieldwork, the strategic examination of archive material can add an objective perspective. Objectivity is introduced by using the research questions to structure how data are collected from archives. In addition, exploring archive material can augment fieldwork by addressing a gap or exhausting an aspect of the research. Archives are not limited to documents or records but include letters, diaries, photographs, videos, audiotapes, artwork, and the like.

## Monograph

Ethnography, in the traditional sense, is a descriptive, interpretive monograph of cultural patterning or social meaning of human social behavior; an ethnography is not an experiential account of fieldwork. The meanings derived from fieldwork constitute the essence of an ethnography. Beyond capturing and conveying a people's worldview, an ethnography should be grounded in social context. The emphasis is at once on specificity and circumstantiality. In analyzing the data of fieldwork and compiling the findings, the ethnographer is challenged to navigate the tension between a positivist perspective (i.e., objectivity) and an interpretivist one (i.e., intellectual abstracting) to achieve "thick description." For an ethnography, raw data are not simply used to describe and itemize findings. Instead, to achieve thick or refined description, a theoretical framework (e.g., discipline, contemporary) is used to describe and conceptualize data.

Another tension challenging the ethnographer in writing a monograph is that of representation. Can the ethnographer understand "others" to the point of speaking for or standing in for them? The irony of ethnography is that the more evolved the interpretation of the data becomes, the more abstract it becomes and the more it becomes one way of representing the data. Clifford Geertz reasoned that ethnography is a "refinement of debate." Juxtaposed to representation as refinement is reflexivity as awareness. Reflexivity is the ethnographer's self-awareness of his or her perspective, political alliance, and cultural influence. The reflexivity process of the ethnographer owning his or her authorial voice and recognizing the voice of others leads to an ongoing deconstructive exercise. This critical exercise maintains an awareness of the multiple realities constructing the reality being represented. Thus, ethnography represents one refined picture of a group of people, not the picture.

*Karen E. Caines*

***See also*** Action Research; Discourse Analysis; Interviewing; Naturalistic Observation; NVivo; Qualitative Research

**Further Readings**

Agar, M. H. (1996). *The professional stranger: An informal introduction to ethnography* (2nd ed.). New York: Academic Press.

Atkinson, P., Coffey, A., Delamont, S., Lofland, J., & Lofland, L. (Eds.). (2001). *Handbook of ethnography.* London: Sage.

Bernard, H. R. (2005). *Research methods in anthropology: Qualitative and quantitative approaches* (4th ed.). Lanham, MD: AltaMira.

Coffey, A. (1999). *The ethnographic self.* London: Sage.

Reed-Danahay, D. E. (Ed.). (1997). *Auto/ethnography: Rewriting the self and the social.* Oxford, UK: Berg.

Van Maanen, J. (Ed.). (1995). *Representation in ethnography.* Thousand Oaks, CA: Sage.

Wolcott, H. F. (2008). *Ethnography: A way of seeing.* Lanham, MD: AltaMira.

# Evidence-Based Decision Making

Quantitative research is a means by which one can gain factual knowledge. Today, a number of such study design options are available to the researcher. However, when studying these quantifications, the researcher must still make a judgment on the interpretation of those statistical findings. To others, what may seem like conflicting, confusing, or ambiguous data requires thoughtful interpretation. Information for decision making is said to be almost never complete, and researchers

always work with a certain level of uncertainty. The need for making and providing proper interpretation of the data findings requires that error-prone humans acknowledge how a decision is made.

Much of the knowledge gained in what is known as the evidence-based movement comes from those in the medical field. Statisticians are an integral part of this paradigm because they aid in producing the evidence through various and appropriate methodologies, but they have yet to define and use terms encompassing the evidence-based mantra. Research methodology continues to advance, and this advancement contributes to a wider base of information. Because of this, the need continues to develop better approaches to the evaluation and utilization of research information. However, such advancements will continue to require that a decision be rendered.

In the clinical sciences, evidence-based decision making is defined as a type of informal decision-making process that combines a clinician's professional expertise coupled with the patient's concerns and evidence gathered from scientific literature to arrive at a diagnosis and treatment recommendation. Milos Jenicek further clarified that evidence-based decision making is the systematic application of the best available evidence to the evaluation of options and to decision making in clinical, management, and policy settings.

Because there is no mutually agreed-upon definition of evidence-based decision making among statisticians, a novel definition is offered here. In statistical research, evidence-based decision making is defined as using the findings from the statistical measures employed and correctly interpreting the results, thereby making a rational conclusion. The evidence that the researcher compiles is viewed scientifically through the use of a defined methodology that values systematic as well as replicable methods for production.

The value of an evidence-based decision-making process provides a more rational, credible basis for the decisions a researcher and/or clinician makes. In the clinical sciences, the value of an evidence-based decision makes patient care more efficient by valuing the role the patient plays in the decision-making process. In statistical research, the value of an evidence-based approach ensures that the methodology used, as well as the logic of the researcher, to arrive at conclusions are sound.

This entry explores the history of the evidence-based movement and the role of decision analysis in evidence-based decision making. In addition, algorithms and decision trees, and their differences, are examined.

## History and Explanation of the Evidence-Based Movement

Evidence-based decision making stems from the evidence-based medicine movement that began in Canada in the late 1980s. David Sackett defined the paradigm of evidence-based medicine/practice as the conscientious, explicit, and judicious use of current best evidence about the care of individual patients. The evidence-based movement has grown rapidly. In 1992, there was one publication on evidence-based practices; by 1998, there were in excess of 1,000. The evidence-based movement continues to enjoy rapid growth in all areas of health care and is seeing headways made in education.

From Sackett's paradigm definition comes a model that encompasses three core pillars, all of which are equally weighted. These three areas are practitioner experience and expertise, evidence from quantitative research, and individual (patient) preferences.

The first pillar in an evidence-based model is the practitioner's individual expertise in his or her respective field. To make such a model work, the individual practitioner has to take into consideration biases, past experience, and training. Typically, the practitioner, be it a field practitioner or a doctoral-level statistician, has undergone some form of mentoring in his or her graduate years. Such a mentorship encourages what is known as the apprentice model, which, in and of itself, is authoritarian and can be argued to be completely subjective. The evidence-based decision-making model attempts to move the researcher to use the latest research findings on statistical methodology instead of relying on the more subjective authoritarian model (mentor–student).

The second pillar in an evidence-based model for the researcher is the use of the latest research

findings that are applicable for the person's field of study. It relies mainly on systematic reviews and meta-analysis studies followed by randomized controlled trials. These types of studies have the highest value in the hierarchy of evidence. Prior to the evidence-based movement in the field of medicine, expert opinion and case studies, coupled with practitioner experience and inspired by their field-based mentor, formed much of the practice of clinical medicine.

The third pillar of an evidence-based model is patient preferences. In the clinical sciences, the need for an active patient as opposed to a passive patient has become paramount. Including the patient in the decision making about his or her own care instills in the patient a more active role. Such an active role by the patient is seen to strengthen the doctor–patient encounter. For the statistician and researcher, it would appear that such a model would not affect their efforts. However, appreciation of this pillar in the scientific research enterprise can be seen in human subject protection.

From this base paradigm arose the term *evidence-based decision making*. Previously, researchers made decisions based on personal observation, intuition, and authority, as well as belief and tradition. Although the researcher examined the evidence that was produced from the statistical formulas used, he or she still relied on personal observation, intuition, authority, belief, and tradition. Interpretation of statistical methods is only as good as the person making the interpretation of the findings.

## Decision Analysis

Decision analysis is the discipline for addressing important decisions in a formal manner. It is composed of the philosophy, theory, methodology, and professional practice to meet this end. John Last suggested that decision analysis is derived from game theory, which tends to identify all available choices and the potential outcomes of each.

The novice researcher and/or statistician may not always know how to interpret results of a statistical test. Moreover, statistical analysis can become more complicated because the inexperienced researcher does not know which test is more suitable for a given situation. Decision analysis is a method that allows one to make such determinations. Jenicek suggested that decision analysis is not a direction-giving method but rather a direction-finding method. Direction-giving methods will be described later on in the form of the decision tree and the algorithm.

Jenicek suggested that decision analysis has seven distinct stages. The first stage of decision analysis requires one to adequately define the problem. The second stage in the decision analysis process is to provide an answer to the question, "What is the question to be answered by decision analysis?" In this stage, true positive, true negative, false negative, and false positive results, as well as other things, need to be taken into consideration. The third stage in the process is the structuring of the problem over time and space. This stage encompasses several key aspects. The researcher must recognize the starting decision point, make an overview of possible decision options and their outcomes, and establish temporo-spatial sequence. Deletion of unrealistic and/or impossible or irrelevant options is also performed at this stage. The fourth stage in the process involves giving dimension to all the relevant components of the problem. This is accomplished by obtaining available data to figure out probabilities. Obtaining the best and most objective data for each relevant outcome is also performed here. The fifth stage is the analysis of the problem. The researcher will need to choose the best way through the available decision paths. As well, the researcher will evaluate the sensitivity of the preferred decision. This stage is marked by the all-important question, "What would happen if conditions of the decision were to change?" The final two stages are solve the problem and act according to the result of the analysis. In evidence-based decision making, the stages that involve the use of the evidence is what highlights this entire process. The decision in evidence-based decision making is hampered if the statistical data are flawed. For the statistician, research methodology using this approach will need to take into consideration efficacy (can it work?), effectiveness (does it work?), and efficiency (what does it cost in terms of time and/or money for what it gives?).

In evidence-based decision-making practice, much criticism has been leveled at what may appear to some as a reliance on statistical measures. It should be duly noted that those who

follow the evidence-based paradigm realize that evidence does not make the decision. However, those in the evidence-based movement acknowledge that valid and reliable evidence is needed to make a good decision.

Jenicek has determined that decision analysis has its own inherent advantages and disadvantages. Advantages to decision analysis are that it is much less costly than the search for the best decision through experimental research. Such experimental research is often sophisticated in design and complex in execution and analysis. Another advantage is that decision analysis can be easily translated into clinical decisions and public health policies. There is also an advantage in the educational realm. Decision analysis is an important tool that allows students to better structure their thinking and to navigate the maze of the decision-making process. A disadvantage to decision analysis is that it can be less valuable if the data and information are of poor quality.

## Algorithm

John Last defined an algorithm as any systematic process that consists of an ordered sequence of steps with each step depending on the outcome of the previous one. It is a term that is commonly used to describe a structured process. It is a graphical representation commonly seen as a flow chart. An algorithm can be described as a specific set of instructions for carrying out a procedure or solving a problem. It usually requires that a particular procedure terminate at some point when questions are readily answered in the affirmative or negative. An algorithm is, by its nature, a set of rules for solving a problem in a finite number of steps. Other names used to describe an algorithm have been method, procedure, and/or technique. In decision analysis, algorithms have also been defined as decision analysis algorithms. They have been argued to be best suited for clinical practice guidelines and for teaching. One of the criticisms of an algorithm is that it can restrict critical thought.

## Decision Tree

A decision tree is a type of decision analysis. Jenicek defined a decision tree as a graphical representation of various options in such things as health, disease evolution, and policy management. The analysis, which involves a decision to be made, leads to the best option. Choices and/or options that are available at each stage in the thinking process have been likened to branches on a tree—a decision tree. The best option could be the most beneficial, most efficacious, and/or most cost-effective choice among the multiple choices to be made. The graphical representation gives the person who will make the decision a method by which to find the best solution among multiple options. Such multiple options can include choices, actions, and possible outcomes, and their corresponding values.

A decision tree is a classifier in the form of a tree structure where a node is encountered. A decision tree can have either a leaf node or a decision node. A leaf node is a point that indicates the value of a target attribute or class of examples. A decision node is a point that specifies some test to be carried out on a single attribute or value. From this, a branch of the tree and/or subtree can represent a possible outcome of a test or scenario. For example, a decision tree can be used to classify a scenario by starting at the root of the tree and moving through it. The movement is temporarily halted when a leaf node, which provides a possible outcome or classification of the instance, is encountered.

Decision trees have several advantages. First of all, decision trees are simple to understand and interpret. After a brief explanation, most people are able to understand the model. Second, decision trees have a value attached to them even if very little hard data support them. Jenicek suggested that important insights can be generated based on experts describing a situation along with its alternative, probabilities, and cost, as well as the experts' preference for a suitable outcome. Third, a decision tree can easily replicate a result with simple math. The final advantage to a decision tree is that it can easily incorporate with other decision techniques. Overall, decision trees represent rules and provide a classification as well as prediction. More important, the decision tree as a decision-making entity allows the researcher the ability to explain and argue why the reason for a decision is crucial. It should be noted that not everything that has branches can be considered a decision tree.

### Decision Tree and the Algorithm: The Differences

A decision tree and an algorithm appear to be quite synonymous and are often confused. Both are graphical representations. However, the difference between the two is that a decision tree is not direction-giving, whereas an algorithm is. In other words, a decision tree provides options to a problem in which each possible decision can be considered and/or argued as pertinent given the situation. An algorithm provides a step-by-step guide in which a decision depends unequivocally on the preceding decision.

### Final Thoughts

Evidence-based decision making, an offshoot of the evidence-based medicine movement, has strong implications for the researcher. Decision analysis should not be confused with direction-giving. Decision analysis is about determination of the finding of direction. Evidence-based decision making takes into consideration that it is a formal method using the best available evidence either from the statistical methodology used in a single study or evidence from multiple studies.

Proponents of evidence-based decision making have always recognized that evidence alone is never the sole determinant of a decision. However, from quantitative research, the evidence from the statistical formulas used, such proponents know that good evidence is needed in order to make a good decision. This is the crux of evidence-based decision making.

*Timothy Mirtz and Leon Greene*

***See also*** Decision Rule; Error; Error Rates; Reliability; Validity of Measurement

#### Further Readings

Friis, R. H., & Sellers, T. A. (1999). *Epidemiology for public health practice* (2nd ed.). Gaithersburg, MD: Aspen.

Jenicek, M. (2003). *Foundations of evidence-based medicine*. Boca Raton, FL: Parthenon.

Jenicek, M., & Hitchcock, D. L. (2005). *Logic and critical thinking in medicine*. Chicago: AMA Press.

Last, J. M. (2001). *A dictionary of epidemiology* (4th ed.). Oxford, UK: Oxford University Press.

Sackett, D. L., Staus, S. E., Richardson, W. S., Rosenberg, W., & Haynes, R. B. (2000). *Evidence-based medicine: How to practice and teach EBM*. Edinburgh, UK: Churchill Livingstone.

## Exclusion Criteria

Exclusion criteria are a set of predefined definitions that is used to identify subjects who will not be included or who will have to withdraw from a research study after being included. Together with inclusion criteria, exclusion criteria make up the eligibility criteria that rule in or out the participants in a research study. Similar to inclusion criteria, exclusion criteria are guided by the scientific objective of the study and have important implications for the scientific rigor of a study as well as for assurance of ethical principles. Commonly used exclusion criteria seek to leave out subjects not complying with follow-up visits, those who are not able to provide biological specimens and data, and those whose safety and ethical protection cannot be assured.

Some definitions are needed to discuss exclusion criteria. Generalizability refers to the applicability of study findings in the sample population to the target population (representativeness) from which the sample was drawn; it requires an unbiased selection of the sample population, which is then said to be generalizable to, or representative of, the target population. Ascertaining exclusion criteria requires screening subjects using valid and reliable measurements to ensure that subjects who are said to meet those criteria really have them (sensitivity) and those who are said not to have them really do not have them (specificity). Such measurements should also be valid (i.e., should truly measure the exclusion criteria) and reliable (consistent and repeatable every time they are measured).

The precision of exclusion criteria will depend on how they are ascertained. For example, ascertaining an exclusion criterion as "self-reported smoking" will likely be less sensitive, specific, valid, and reliable than ascertaining it by means of testing for levels of cotinine in blood. On the other

hand, cotinine in blood may measure exposure to secondhand smoking, thus excluding subjects who should not be excluded; therefore, a combination of self-reported smoking and cotinine in blood may increase the sensitivity, specificity, validity, and reliability of such measurement, but it will be more costly and time consuming.

A definition of exclusion criteria that requires several measurements may be just as good as one using fewer measurements. Good validity and reliability of exclusion criteria will help minimize random error, selection bias, and confounding, thus improving the likelihood of finding an association, if there is one, between the exposures or interventions and the outcomes; it will also decrease the required sample size and allow representativeness of the sample population. Using standardized exclusion criteria is necessary to accomplish consistency, replicability, and comparability of findings across similar studies on a research topic. Standardized disease-scoring definitions are available for mental and general diseases (*Diagnostic and Statistical Manual of Mental Disorders* and *International Classification of Diseases,* respectively). Study results on a given research topic should carefully compare the exclusion criteria to analyze consistency of findings and applicability to sample and target populations. Exclusion criteria must be as parsimonious in number as possible; each additional exclusion criterion may decrease sample size and result in selection bias, thus affecting the internal validity of a study and the external validity (generalizability) of results, in addition to increasing the cost, time, and complexity of recruiting study participants. Exclusion criteria must be selected carefully based upon a review of the literature on the research topic, in-depth knowledge of the theoretical framework, and their feasibility and logistic applicability.

Research proposals submitted for institutional review board (IRB) approval should clearly describe exclusion criteria to potential study participants, as well as consequences, at the time of obtaining informed consent. Often, research protocol amendments that change the exclusion criteria will result in two different sample populations that may require separate data analyses with a justification for drawing composite inferences. Exceptions to exclusion criteria need to be approved by the IRB of the research institution; changes to exclusion criteria after IRB approval require new approval of any amendments.

In epidemiologic and clinical research, assessing an exposure or intervention under strict study conditions is called *efficacy,* whereas doing so in real-world settings is called *effectiveness*. Concerns have been raised about the ability to generalize the results from randomized clinical trials to a broader population, because participants are often not representative of those seen in clinical practice. Each additional exclusion criterion implies a different sample population and approaches the assessment of efficacy, rather than effectiveness, of the exposure or intervention under study, thus influencing the utility and applicability of study findings. For example, studies of treatment for alcohol abuse have shown that applying stringent exclusion criteria used in research settings to the population at large results in a disproportionate exclusion of African Americans, subjects with low socioeconomic status, and subjects with multiple substance abuse and psychiatric problems. Therefore, the use of more permissive exclusion criteria has been recommended for research studies on this topic so that results are applicable to broader, real-life populations. The selection and application of these exclusion criteria will also have important consequences on the assurance of ethical principles, because excluding subjects based on race, gender, socioeconomic status, age, or clinical characteristics may imply an uneven distribution of benefits and harms, disregard for the autonomy of subjects, and lack of respect. Researchers must strike a balance between stringent and more permissive exclusion criteria. On one hand, stringent exclusion criteria may reduce the generalizability of sample study findings to the target population, as well as hinder recruitment and sampling of study subjects. On the other hand, they will allow rigorous study conditions that will increase the homogeneity of the sample population, thus minimizing confounding and increasing the likelihood of finding a true association between exposure/intervention and outcomes. Confounding may result from the effects of concomitant medical conditions, use of medications other than the one under study, surgical or rehabilitation interventions, or changes in the severity of

disease in the intervention group or occurrence of disease in the nonintervention group.

Changes in the baseline characteristics of study subjects that will likely affect the outcomes of the study may also be stated as exclusion criteria. For example, women who need to undergo a specimen collection procedure involving repeated vaginal exams may be excluded if they get pregnant during the course of the study. In clinical trials, exclusion criteria identify subjects with an unacceptable risk of taking a given therapy or even a placebo (for example, subjects allergic to the placebo substance). Also, exclusion criteria will serve as the basis for contraindications to receive treatment (subjects with comorbidity or allergic reactions, pregnant women, children, etc.). Unnecessary exclusion criteria will result in withholding treatment from patients who may likely benefit from a given therapy and preclude the translation of research results into practice. Unexpected reasons for subjects' withdrawal or attrition after inception of the study are not exclusion criteria. An additional objective of exclusion criteria in clinical trials is enhancing the differences in effect between a drug and a placebo; to this end, subjects with short duration of the disease episode, those with mild severity of illness, and those who have a positive response to a placebo may be excluded from the study.

*Eduardo Velasco*

***See also*** Bias; Confounding; Inclusion Criteria; Reliability; Sampling; Selection; Validity of Measurement; Validity of Research Conclusions

**Further Readings**

Gordis, L. (2008). *Epidemiology* (4th ed.). Philadelphia: W. B. Saunders.

Hulley, S. B., Cummings, S. R., Browner, W. S., Grady, D., & Newman, T. B. (Eds.). (2007). *Designing clinical research: An epidemiologic approach* (3rd ed.). Philadelphia: Lippincott Williams & Wilkins.

LoBiondo-Wood, G., & Haber, J. (2006). *Nursing research: Methods and critical appraisal for evidence-based practice* (6th ed.). St. Louis, MO: Mosby.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Szklo, M., & Nieto, F. J. (2007). *Epidemiology: Beyond the basics*. Boston: Jones & Bartlett.

# Exogenous Variables

*Exogenous* originated from the Greek words *exo* (meaning "outside") and *gen* (meaning "born"), and describes something generated from outside a system. It is the opposite of endogenous, which describes something generated from within the system. Exogenous variables, therefore, are variables that are not caused by any other variables in a model of interest; in other words, their value is not determined in the system being studied.

The concept of exogeneity is used in many fields, such as biology (an exogenous factor is a factor derived or developed from outside the body); geography (an exogenous process takes place outside the surface of the earth, such as weathering, erosion, and sedimentation); and economics (exogenous change is a change coming from outside the economics model, such as changes in customers' tastes or income for a supply-and-demand model). Exogeneity has both statistical and causal interpretations in social sciences. The following discussion focuses on the causal interpretation of exogeneity.

## Exogenous Variables in a System

Although exogenous variables are not caused by any other variables in a model of interest, they may cause the change of other variables in the model. In the specification of a model, exogenous variables are usually labeled with *X*s and endogenous variables are usually labeled with *Y*s. Exogenous variables are the "input" of the model, predetermined or "given" to the model. They are also called *predictors* or *independent* variables.

The following is an example from educational research. Family income is an exogenous variable to the causal system consisting of preschool attendance and student performance in elementary school. Because family income is determined by neither a student's preschool attendance nor elementary school performance, family income is an exogenous variable to the system being studied. On the other hand, students' family income may determine both preschool attendance and

elementary school performance. High-income families are more likely than low-income families to enroll their children in preschools. High-income families also tend to provide more resources and support for their children to perform well in elementary school; for example, high-income parents may purchase more learning materials and spend more spare time helping their children with homework assignments than low-income parents.

Whether a variable is exogenous is relative. An exogenous variable in System A may not be an exogenous variable in System B. For example, family income is an exogenous variable in the system consisting of preschool attendance and elementary school performance. However, family income is not an exogenous variable in the system consisting of parental education level and parental occupation because parental education level and occupation probably influence family income. Therefore, once parental education level and parental occupation are added to the system consisting of family income, family income will become an endogenous variable.

## Exogenous Variables in Path Analysis

Exogenous and endogenous variables are frequently used in structural equation modeling, especially in path analysis in which a path diagram can be used to portray the hypothesized causal and correlational relationships among all the variables. By convention, a hypothesized causal path is indicated by a single-headed arrow, starting with a cause and pointing to an effect. Because exogenous variables do not receive causal inputs from other variables in the system, no single-headed arrow points to exogenous variables. Figure 1 illustrates a hypothetical path diagram.

In this model, family income is an exogenous variable. It is not caused by any other variables in the model, so no single-headed arrow points to it. A child's preschool attendance and reading/math scores are endogenous variables. Each of the endogenous variables is influenced by at least one other variable in the model and accordingly has at least one single-headed arrow pointing to it. Family income is hypothesized to produce changes in all the other variables; therefore, single-headed arrows start from family income and end with the child's preschool attendance, reading score, and math score, respectively. Moreover, a child's preschool attendance is hypothesized to influence the child's reading and math scores; thus, single-headed arrows start from the child's preschool attendance and end with the child's reading and math scores. Finally, a child's reading score is hypothesized to influence the child's math score; therefore, a single-headed arrow points to the math score from the reading score. In this model, even though a child's preschool attendance and reading score both cause changes in other variables, they are not exogenous variables because they both have input from other variables as well. As preschool attendance and reading score serve as independent variables as well as dependent variables, they are also called mediators.

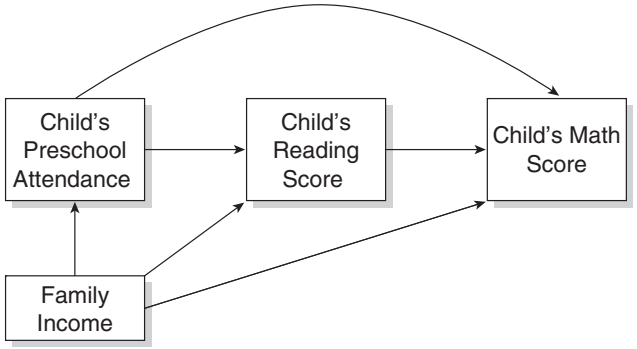



**Figure 1** A Hypothetical Path Diagram

More than one exogenous variable may exist in a path analysis model, and the exogenous variables in a model may be correlated with each other. A correlation relationship is conventionally indicated by a double-headed arrow in path analysis. Therefore, in a path diagram, two exogenous variables may be connected with a double-headed arrow.

Finally, it is worth mentioning that the causal relationships among exogenous and endogenous variables in a path analysis may be hypothetical and built on theories and/or common sense. Therefore, researchers should be cautious and consider research design in their decision about whether the hypothesized causal relationships truly exist, even if the hypothesized causal relationships are supported by statistical results.

*Yue Yin*

*See also* Cause and Effect; Endogenous Variables; Path Analysis

### Further Readings

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford.
Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge, UK: Cambridge University Press.
Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. New York: Houghton Mifflin.

## Expected Value

The expected value is the mean of all values of a random variable weighted by the probability of the occurrence of the values. The expected value (or expectation, or mean) of random variable (RV) $X$ is denoted as $E[X]$ (or sometimes $\mu$).

### Mathematical Definition

The RV, $X$, of a random experiment, which is defined on a probability space $(\Omega, \Sigma, P)$ on an underlying sample space $\Omega$, takes value in event set $\Sigma \subseteq \mathbb{R}$ with certain probability measure $P$.

If $X$ is a continuous random variable (i.e., $\Sigma$ is an interval), the expected value of $X$ is defined as

$$E[X] = \int_{\Omega} x dP.$$

If $X$ also has a probability density function (pdf) $f(x)$ of certain probability distribution, the above expected value of $X$ can be formulated as

$$E[X] = \int_{x \in \Omega} x \cdot f(x) dx.$$

The expected value exists if the above integral of absolute value of $X$ is absolutely convergent, that is, $\int_{x \in \Omega} |x| dx$ is finite.

If $X$ is a discrete random variable (i.e., $\Omega$ is countable) with probability mass or probability density function (pmf or pdf) $p(x)$, the expected value of $X$ is defined as

$$E[x] = \sum_{x \in \Omega} x \cdot p(x),$$

and the above expected value exists if the above sum of absolute value of $X$ is absolutely convergent, that is, $\sum_{x \in \Omega} |x|$ is finite.

As a simple example, suppose $x$ can take on two values, 0 and 1, which occur with probabilities .4 and .6. Then

$$E[X] = (x = 0) \times p(x = 0) + (x = 1) \times p(x = 1) = 0 \times 0.4 + 1 \times 0.6 = 0.6.$$

If $G(X)$ is a function of RV $X$, its expected value, $E[G(X)]$, is a weighted average of the possible values of $G(X)$ and is defined as

$$E[G(X)] = \begin{cases} \int_{x \in \Omega} G(x) \cdot f(x) dx, \text{for continuous case} \\ \sum_{x \in \Omega} G(x) \cdot p(x), \text{for discrete case} \end{cases}.$$

The above expected value exists if the above sum of absolute value of $G(X)$ is absolutely convergent.

### Linear Relationship

If RVs $X_1, X_2, \ldots, X_n$ have the expectations as $\mu_1, \mu_2, \ldots, \mu_n$, and $c_1, c_2, \ldots, c_n$ are all constants, then

$$\therefore E[c_0] = c_0 \text{ and } E[c_i X_i] = c_i E[X_i] = c_i \mu_i \quad (1)$$

$$\because E[c_i X_i + c_0] = c_i E[X_i] + c_0 = c_i \mu_i + c_0 \quad (2)$$

$$E\left[\sum_{i=1}^{n} c_i X_i + c_0\right] = \sum_{i=1}^{n} c_i E[X_i] + c_0 = \sum_{i=1}^{n} c_i \mu_i + c_0 \quad (3)$$

$$E\left[\sum_{i=1}^{n} c_i G_i(X_i) + c_0\right] = \sum_{i=1}^{n} c_i E[G_i(X_i)] + c_0 = \sum_{i=1}^{n} c_i \sum_{x \in \Omega} G_i(x) \cdot p(x) + c_0 \quad (4)$$

$$E[X_1 X_2 \ldots X_n] = \mu_1 \mu_2 \ldots \mu_n \text{ for independent } X_i. \quad (5)$$

### Interpretation

From a statistical point of view, the following terms are important: arithmetic mean (or simply mean), central tendency, and location statistic. The arithmetic mean of $X$ is the summation of the set of observations (sample) of $X = \{x_1, x_2, \ldots, x_N\}$ divided by the sample size $N$:

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} x_i.$$

$\overline{X}$ is called arithmetic mean/sample mean/average when used to estimate the location of a sample. When it is used to estimate the location of an underlying distribution, $\overline{X}$ is called population mean/average, or expectation/expected value, which can be denoted as $E[X]$ or $\mu$. This is consistent with the original definition because the probability of each value's occurrence is equal to $1/N$. One could construct a different estimate of the mean, for example, if some values were expected to occur more frequently than others. Because a researcher rarely has such information, the simple mean is most commonly used.

### Moment

The moment (a characteristic of a distribution) of $X$ about the real number $c$ is defined as

$$E[(x-c)^n] \forall c \in \mathbb{R}, \text{ and integer } n \geq 1.$$

Hence, $E[X^n]$ are also called central moments. $E[X]$ is called the first moment ($\because n = 1$) of $X$ about $c = 0$, which is commonly called the mean of $X$. *The second moment about the mean of X* is called the Variance of $X$. Theoretically, the entire distribution of $X$ can be described if all moments of $X$ are known by using the moment-generating functions, although only the first five moments are generally necessary to specify a distribution completely. The third moment is termed *skewness,* and the fourth is called *kurtosis*.

### Joint Expected Value

If $X_i$,$X_j$ are continuous random variables with the joint pdf $f(x_i, x_j)$, the expected value of $G(X_i,X_j)$ is

$$E\left[G\left(X_i,X_j\right)\right] = \int_{x_i \in \Omega_{x_i}} \int_{x_j \in \Omega_{x_j}} G\left(x_i,x_j\right) f\left(x_i,x_j\right) dx_i dx_j,$$

$\forall i,j \in \mathbb{N}$, and $i \neq j$.

If $X_i$,$X_j$ are discrete random variables with the joint pmf $p(x_i,x_j)$, the expected value of $G(X_i,X_j)$ is

$$E\left[G\left(X_i,X_j\right)\right] = \sum_{x_i \in \Omega_{x_i}} \sum_{x_j \in \Omega_{x_j}} G\left(x_i,x_j\right) p\left(x_i,x_j\right) \ \forall i,j \in \mathbb{N},$$

and $i \neq j$.

### Conditional Expected Value

Given that $X_i$,$X_j$ are continuous random variables with the joint pdf $f(x_i,x_j)$ and $f(x_j) > 0$, then the conditional pdf of $X_i$ given $X_j = x_j$ is

$$f_{X_i|X_j}\left(x_i|x_j\right) = \frac{f\left(x_i,x_j\right)}{f_{X_j}\left(x_j\right)}, \forall \text{ all } x_i,$$

and the corresponding conditional expected value is

$$E\left[X_i|X_j = x_j\right] = \int_{x_i \in \Omega_{x_i}} x_i \cdot f_{X_i|X_j}\left(x_i|x_j\right) dx_i.$$

Given that $X_i$,$X_j$ are discrete random variables with the joint pmf $p(x_i,x_j)$ and $p(x_j) > 0$, then the conditional pdf of $X_i$ given $X_j = x_j$ is

$$p_{X_i|X_j}\left(x_i|x_j\right) = \frac{p\left(x_i,x_j\right)}{p_{X_j}\left(x_j\right)}, \forall \text{ all } x_i,$$

and the corresponding conditional expected value is

$$E\left[X_i|X_j = x_j\right] = \sum_{x_i \in \Omega_{x_i}} x_i \cdot p_{X_i|X_j}\left(x_i|x_j\right).$$

Furthermore, the expectation of the conditional expectation of $X_i$ given $X_j = x_j$ is simply the expectation of $X_i$:

$$E\left[E\left[X_i|X_j\right]\right] = E[X_i].$$

## Variance and Covariance

Given a continuous RV $X$ with pdf $f(x)$, the variance of $X$ can be formulated as

$$Var[X] = \sigma_X^2 = E\left[(X-\mu)^2\right] = \int_{x\in\Omega}(x-\mu)^2 \cdot f(x)\,dx.$$

When $X$ is a discrete RV with pmf $p(x)$, the variance of $X$ can be formulated as

$$Var[X] = \sigma_X^2 = E\left[(X-\mu)^2\right] = \sum_{x\in\Omega}(x-\mu)^2 \cdot p(x).$$

The positive square root of variance, $\left|\sqrt{\sigma_X^2}\right|$, is called *standard deviation* and denoted as $\sigma_X$ or $s_X$.

Given two continuous RVs $X$ and $Y$ with joint pdf $f(x,y)$, the covariance of $X$ and $Y$ can be formulated as

$$Cov[X,Y] = \sigma_{XY} = E[(X-\mu_X)(Y-\mu_Y)] = \int_{x\in\Omega_X}\int_{y\in\Omega_Y}(x-\mu_X)(y-\mu_Y)\cdot f(x,y)\,dx.$$

When $X$ and $Y$ are discrete RVs with joint pmf $p(x,y)$, the covariance of $X$ and $Y$ can be formulated as

$$Cov[X,Y] = \sigma_{XY} = E[(X-\mu_X)(Y-\mu_Y)] = \sum_{x\in\Omega_X}\sum_{y\in\Omega_Y}(x-\mu_X)(y-\mu_Y)\cdot p(x,y).$$

### *Linear Relationship of Variance and Covariance*

If $X$, $Y$, and $Z$ are random variables and $c_0$, $c_1, \ldots, c_n$ are all constants, then

$$Var[X] = E\left[X^2\right] - E[|X|]^2 = Cov[X,X] \quad (6)$$

$$\because Var[c_0] = 0 \text{ and } Var[c_iX] = c_i^2 Var[X] \quad (7)$$

$$\therefore Var[c_iX + c_0] = c_i^2 Var[X] = c_i^2\sigma_X^2 \quad (8)$$

$$Var\left[c_iX + c_jY\right] = c_i^2 Var[X] + c_j^2 Var[Y] + 2c_ic_jCov[X,Y] \quad (9)$$

$$Cov[X,Y] = E[XY] - E[X]E[Y] = Cov[X,Y] \quad (10)$$

$$Cov\left[c_iX + c_jY, Z\right] = c_iCov[XZ] + c_jCov[YZ]. \quad (11)$$

## Uncorrelatedness Versus Independence

Continuous random variables $X_i$ and $X_j$ are said to be *independent* if and only if the joint pdf (or joint pmf of discrete RVs) of $X_i$ and $X_j$ equals the product of the marginal pdfs for $X_i$ and $X_j$, respectively:

$$f(X_i, X_j) = f(X_i)f(X_j).$$

When $X_i$, $X_j$ are said to be independent, the conditional expected value of $X_i$ given $X_j = x_j$ is the expectation of $X_i$:

$$E\left[X_i|X_j\right] = E[X_i]. \quad (12)$$

Also, the expectation of the product of $X_i$ and $X_j$ equals the product of the expectations for $X_i$ and $X_j$, respectively:

$$E\left[X_iX_j\right] = E[X_i]E\left[X_j\right]. \quad (13)$$

$X_i$ and $X_j$ are said to be *uncorrelated*, *orthogonal*, and *linearly independent* if their covariance is zero, that is, $Cov[X_i,X_j] = 0$. Then the variance of the summation of $X_i$ and $X_j$ equals the summation of the variances of $X_i$ and $X_j$, respectively. That is, Equation 9 can be rewritten as

$$Var\left[X_i, X_j\right] = Var[X_i] + Var\left[X_j\right]. \quad (14)$$

Independence is a stronger condition that always implies uncorrelatedness and orthogonality, but not vice versa. For example, a perfect spherical relationship between $X$ and $Y$ will be uncorrelated but certainly not independent. Because simple correlation is linear, many relationships would have a zero correlation but be interpretably nonindependent.

## Inequalities

### *Basic Inequality*

Given $X_i$,$X_j$ are random variables, if the realization $x_i$ is always less than or equal to $x_j$, the

expected value of $X_i$ is less than or equal to that of $X_j$:

$$\because X_i \le X_j, \therefore E[X_j] \le E[Y_i]. \tag{15}$$

The expected value of the absolute value of a random variable $X$ is less than or equal to the absolute value of its expectation:

$$E[|X|] \ge |E[X]|. \tag{15}$$

### *Jensen Inequality*

For a convex function $h(\cdot)$ and RV $X$ with pdf/pmf $p(x)$, then

$$E[h(X)] \ge h(E[X]). \tag{16}$$

Or, put in another way,

$$\sum_{x \in \Omega_X} h(x) \cdot p(x) \ge h\left(\sum_{x \in \Omega_X} x \cdot p(x)\right). \tag{17}$$

### *Markov Inequality*

If $X \ge 0$, then for all $x \ge 0$, the Markov inequality is defined as

$$P(X \ge x) \le \frac{E[X]}{x}. \tag{18}$$

### *Chebyshev Inequality*

If the variance of $X$ is known, the tighter Chebyshev inequality bound is

$$P(|X - E(X)| \ge x) \le \frac{Var[X]}{x^2}. \tag{19}$$

## Expectations and Variances for Well-Known Distributions

The following table lists distribution characteristics, including the expected value of $X$ and the expected variance of both discrete and continuous variables.

## Estimation

Given the real-valued random variables $Y$ and $X$, the optimal minimum mean-square error (MMSE) estimator of $Y$ based on observing $X$ is the conditional mean, $E[Y|X]$, which in the conditional expectation across the ensemble of all random process with the same and finite second moment. Also, the MMSE $\varepsilon$ of $Y$ given $X$, which is defined as $\varepsilon = Y - E[Y|X]$, is orthogonal to any function of data $X$, $G(X)$. That is the so-called orthogonality principle

$$E[\varepsilon \cdot G(X)] = 0.$$

When the observations $Y$ are normally distributed with zero mean, the MMSE estimator of $Y$ given $X$, the conditional expectation $E[Y|X]$, is linear. For those non-normally distributed $Y$, the conditional expectation estimator can be nonlinear. Usually, it is difficult to have the close form of the optimal nonlinear estimates, which will depend on higher order moment functions.

Given a given length of data set $(Y, X) = \{(y_i, x_i);\ i = 1, 2, \ldots, N\}$ with sample size $N$, the linear least square error (LSE) estimator of $Y$ given $X$ is

$$\hat{Y} = \hat{b}_1 X_1 + \hat{b}_0,$$

where

| *Discrete Random Variables* | | | *Continuous Random Variables* | | |
|---|---|---|---|---|---|
| Distribution | $E[X]$ | $Var[X]$ | Distribution | $E[X]$ | $Var[X]$ |
| Bernoulli ($p$) | $p$ | $p(1-p)$ | Uniform ($a$,$b$) | $(a+b)/2$ | $(b-a)^2/2$ |
| Binomial ($n$,$p$) | $np$ | $np(1-p)$ | Exponential ($\lambda$) | $1/\lambda$ | $1/\lambda^2$ |
| Poisson ($\lambda$) | $\lambda$ | $\lambda$ | Normal ($\mu$,$\sigma^2$) | $\mu$ | $\sigma^2$ |
| Negative Binomial ($r$,$p$) | $r/p$ | $r(1-p)/p^2$ | Gamma ($\alpha$,$\lambda$) | $\alpha/\lambda$ | $\alpha/\lambda^2$ |
| Geometric ($k$,$p$) | $(1-p)/p$ | $(1-p)/p^2$ | Cauchy ($\alpha$) | Undefined | Undefined |

$$\hat{b}_1 = \frac{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \frac{1}{N}\sum_{i=1}^{N} y_i\right)\left(x_i - \frac{1}{N}\sum_{i=1}^{N} x_i\right)}{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \frac{1}{N}\sum_{i=1}^{N} x_i\right)^2}$$

and

$$\hat{b}_0 = \frac{1}{N}\sum_{i=1}^{N} y_i - \hat{b}_1 \cdot \frac{1}{N}\sum_{i=1}^{N} x_i.$$

If the random process is stationary (i.e., constant mean and time shift-invariant covariance) and ergodic (i.e., time average $\frac{1}{N}\sum_{i=1}^{N} x_i$ converges in mean-square sense to the ensemble average $E[X]$), the LSE estimator will be asymptotically equal to the MMSE estimator.

MMSE estimation is not the only possibility for the expected value. Even in simple data sets, other statistics, such as the median or mode, can be used to estimate the expected value. They have properties, however, that make them suboptimal and inefficient under many situations. Other estimation approaches include the mean of a Bayesian posterior distribution, the maximum likelihood estimator, or the biased estimator in ridge regression. In modern statistics, these alternative estimation methods are being increasingly used. For a normally distributed random variable, all will typically yield the same value except for computational variation in computer routines, usually very small in magnitude. For non-normal distributions, various other considerations come into play, such as the type of distribution encountered, amount of data available, and existence of the various moments.

*Victor L. Willson and Jiun-Yu Wu*

***See also*** Chi-Square Test; Sampling Distributions; Sampling Error

**Further Readings**

Brennan, A., Kharroubi, S., O'Hagan, A., & Chilcott, J. (2007). Calculating partial expected value of perfect information via Monte Carlo sampling algorithms. *Medical Decision Making, 27,* 448–470.

Felli, J. C., & Hazen, G. B. (1998). Sensitivity analysis and the expected value of perfect information. *Medical Decision Making, 18,* 95–109.

Glass, G. (1964). [Review of the book *Expected values of discrete random variables and elementary statistics* by A. L. Edwards]. *Educational and Psychological Measurement, 24,* 969–971.

# Experience Sampling Method

The experience sampling method (ESM) is a strategy for gathering information from individuals about their experience of daily life as it occurs. The method can be used to gather both qualitative and quantitative data, with questions for participants that are tailored to the purpose of the research. It is a phenomenological approach, meaning that the individual's own thoughts, perceptions of events, and allocation of attention are the primary objects of study. In the prototypical application, participants in an ESM study are asked to carry with them for 1 week a signaling device such as an alarm wristwatch or palmtop computer and a recording device such as a booklet of questionnaires. Participants are then signaled randomly 5 to 10 times daily, and at each signal, they complete a questionnaire. Items elicit information regarding the participants' location at the moment of the signal, as well as their activities, thoughts, social context, mood, cognitive efficiency, and motivation. Researchers have used ESM to study the effects of television viewing on mood and motivation, the dynamics of family relations, the development of adolescents, the experience of engaged enjoyment (or flow), and many mental and physical health issues. Other terms for ESM include *time sampling, ambulatory assessment,* and *ecological momentary assessment;* these terms may or may not signify the addition of other types of measures, such as physiological markers, to the protocol.

## Designing a Study Using ESM

In ESM studies, researchers need to select a sample of people from a population, but they must also choose a method to select a sample of moments from the population of all moments of experience. Many studies make use of signal-contingent sampling, a stratified random approach in which the day is divided into equal segments and the

participant is signaled at a random moment during each segment. Other possibilities are to signal the participant at the same times every day (interval-contingent sampling) or to ask the participant to respond after every occurrence of a particular event of interest (event-contingent sampling). The number of times per day and the number of days that participants are signaled are parameters that can be tailored based on the research purpose and practical matters.

Increasingly, researchers are using palmtop computers as both the signaling device and the recording device. The advantages here are the direct electronic entry of the data, the ability to time-stamp each response, and the ease of programming a signaling schedule. Disadvantages include the difficulty in obtaining open-ended responses and the high cost of the devices. When a wristwatch or pager is used as the signaling device and a pen with a booklet of blank questionnaires serves as the recording device, participants can be asked open-ended questions such as "What are you doing?" rather than be forced to choose among a list of activity categories. This method is less costly, but does require more coding and data entry labor. Technology appears to be advancing to the point where an inexpensive electronic device will emerge that will allow the entry of open-ended responses with ease, perhaps like text-messaging on a mobile phone.

## Analysis of ESM Data

Data resulting from an ESM study are complex, including many repeated responses to each question. Responses from single items are also often combined to form multi-item scales to measure constructs such as mood or intrinsic motivation. Descriptive information, such as means and frequencies, can be computed at the response level, meaning that each response is treated as one case in the data. However, it is also useful to aggregate the data by computing means within each person and percentages of responses falling in categories of interest (e.g., when with friends). Often, *z*-scored variables standardized to each person's own mean and standard deviation are computed to get a sense of how individuals' experiences in one context differ from their average levels of experiential quality. To avoid the problem of nonindependence in the data, person-level variables are preferred when using inferential statistical techniques such as analysis of variance or multiple regression. More complex procedures, such as hierarchical linear modeling, multilevel modeling, or mixed-effects random regression analysis, allow the researcher to consider the response-level and person-level effects simultaneously.

## Studies Involving ESM

Mihaly Csikszentmihalyi was a pioneer of the method when he used pagers in the 1970s to study a state of optimal experience he called *flow.* Csikszentmihalyi and his students found that when people experienced a high level of both challenges and skills simultaneously, they also frequently had high levels of enjoyment, concentration, engagement, and intrinsic motivation. To study adolescents' family relationships, Reed Larson and Maryse Richards signaled adolescents and their parents simultaneously. The title of their book, *Divergent Realities,* telegraphs one of their primary conclusions. Several researchers have used ESM to study patients with mental illness, with many finding that symptoms worsened when people were alone with nothing to do. Two paradoxes exposed by ESM research are that people tend to retrospectively view their work as more negative and TV-watching as more positive experiences than what they actually report when signaled in the moment while doing these activities.

*Joel M. Hektner*

***See also*** Ecological Validity; Hierarchical Linear Modeling; Levels of Measurement; Multilevel Modeling; Standardized Score; *z* Score

### Further Readings

Bolger, N., Davis, A., & Rafaeli, E. (2003). Diary methods: Capturing life as it is lived. *Annual Review of Psychology, 54,* 579–616.

Hektner, J. M., Schmidt, J. A., & Csikszentmihalyi, M. (2007). *Experience sampling method: Measuring the quality of everyday life.* Thousand Oaks, CA: Sage.

Reis, H. T., & Gable, S. L. (2000). Event-sampling and other methods for studying everyday experience. In H. T. Reis & C. M. Judd (Eds.), *Handbook of*

*research methods in social and personality psychology* (pp. 190–222). New York: Cambridge University Press.
Shiffman, S. (2000). Real-time self-report of momentary states in the natural environment: Computerized ecological momentary assessment. In A. A. Stone, J. S. Turkkan, C. A. Bachrach, J. B. Jobe, H. S. Kurtzman, & V. S. Cain (Eds.), *The science of self-report: Implications for research and practice* (pp. 276–293). Mahwah, NJ: Lawrence Erlbaum.
Walls, T. A., & Schafer, J. L. (Eds.). (2006). *Models for intensive longitudinal data*. New York: Oxford University Press.

# Experimental Design

Empirical research involves an experiment in which data are collected in two or more conditions that are identical in all aspects but one. A blueprint for such an exercise is an experimental design. Shown in Table 1 is the design of the basic experiment. It has (a) one independent variable (*color*) with two levels (pink and white); (b) four control variables (*age, health, sex,* and *IQ*); (c) a control procedure (i.e., random assignment of subjects); and (d) a dependent variable (*affective score*).

## Method of Difference and Experimental Control

Table 1 also illustrates the inductive rule, method of difference, which underlies the basic one-factor, two-level experiment. As *age* is being held constant, any slight difference in *age* between subjects in the two conditions cannot explain the difference (or its absence) between the mean performances of the two conditions. That is, as a control variable, *age* excludes itself from being an explanation of the data.

There are numerous extraneous variables, any one of which may potentially be an explanation of the data. Ambiguity of this sort is minimized with appropriate control procedures, an example of which is random assignment of subjects to the two conditions. The assumption is that, in the long run, effects of unsuspected confounding variables may be balanced between the two conditions.

## Genres of Experimental Designs for Data Analysis Purposes

Found in Column I of Table 2 are three groups of designs defined in terms of the number of factors used in the experiment, namely, one-factor, two-factor, and multifactor designs.

### *One-Factor Designs*

It is necessary to distinguish between the two-level and multilevel versions of the one-factor design because different statistical procedures are used to analyze their data. Specifically, data from a one-factor, two-level design are analyzed with the *t* test. The statistical question is whether or not the difference between the means of the two conditions can be explained by chance influences (see Row *a* of Table 2).

Some version of one-way analysis of variance would have to be used when there are three or more levels to the independent variable (see Row *b* of Table 2). The statistical question is whether or not the variance based on three or more test conditions is larger than that based on chance.

With quantitative factors (e.g., dosage) as opposed to qualitative factors (e.g., type of drug),

**Table 1** Basic Structure of an Experiment

| | | *Control Variables* | | | | *Control Procedure* | |
|---|---|---|---|---|---|---|---|
| *Test Condition* | *Independent Variable Manipulated, Wall Color* | *Age* | *Health* | *Sex* | *IQ* | *Random Assignment of Subjects* ($S_i$) | *Dependent Variable, Affective Score* |
| Experimental | Pink | Middle-aged | Good | Male | Normal | $S_1, S_{21}, S_7, \ldots S_{15}$ | To be collected and analyzed |
| Control | White | Middle-aged | Good | Male | Normal | $S_9, S_{10}, S_{24}, \ldots S_2$ | |

**Table 2** Genres of Experimental Designs in Terms of Treatment Combinations

*Panel A: One-Factor Designs in Terms of Number of Levels*

| | *I* | *II* | *III* | |
|---|---|---|---|---|
| | *Number of Factors* | *Number of Levels in Factor* | *Statistical Test (parametric)* | *Statistical Question* |
| *a* | 1 | 2 | *t* test | Is the difference between the two means accountable by chance influences? |
| *b* | | 3 or more | One-way ANOVA | Can the variance based on the means of the 3 (or more) conditions be explained in terms of chance influences? Are there trends in the data? |
| *c* | 2(*A*, *B*) | $m \times n$ | Two-way ANOVA | **Main effect** of *A*: Is the difference between the *m* means of A accountable by chance influences?<br>**Main effect** of *B*: Is the difference between the *n* means of B accountable by chance influences?<br>AB **interaction**: Can the difference among the means of the $m \times n$ treatment combinations accountable by chance influences?<br>Simple effect of *A*: Is the difference among the *m* means of A at Level *j* of *B* accountable by chance influences? influences?<br>Simple effect of *B*: Is the difference among the *n* means of B at Level *i* of *A* accountable by chance influences? |
| *d* | 3 or more | $m \times n \times p$ or $m \times n \times p \times q$ | Multi-factor ANOVA | Extension of the questions found in two-way ANOVA |

one may ascertain trends in the data when a factor has three or more levels (see Row *b*). Specifically, a minimum of three levels is required for ascertaining a linear trend, and a minimum of four levels for a quadratic trend.

### *Two-Factor Designs*

Suppose that Factors *A* (e.g., room color) and *B* (e.g., room size) are used together in an experiment. Factor *A* has *m* levels; its two levels are $a_1$ and $a_2$ when $m = 2$. If Factor *B* has *n* levels (and if $n = 2$), the two levels of *B* are $b_1$ and $b_2$. The experiment has a factorial design when every level of *A* is combined with every level of *B* to define a test condition or *treatment combination*. The size of the factorial design is *m* by *n*; it has *m*-by-*n* treatment combinations. This notation may be generalized to reflect factorial design of any size.

Specifically, the number of integers in the name of the design indicates the number of independent variables, whereas the identities of the integers stand for the respective number of levels. For example, the name of a three-factor design is *m* by *n* by *p*; the first independent variable has *m* levels, the second has *n* levels, and the third has *p* levels (see Row *d* of Table 2).

The lone statistical question of a one-factor, two-level design (see Row *a* of Table 2) is asked separately for Factors *A* and *B* in the case of the two-factor design (see [a] and [b] in Row *c* of Table 2). Either of them is a main effect (see [a] and [b] in Row *c*) so as to distinguish it from a simple effect (see Row *c*). This distinction may be illustrated with Table 3.

#### Main Effect

Assume an equal number of subjects in all treatment combinations. The means of $a_1$ and $a_2$ are 4.5 and 2.5, respectively (see the "Mean of $a_i$" column in either panel of Table 3). The main effect of *A* is 2 (i.e., $4.5 - 2.5$). In the same vein, the means of $b_1$ and $b_2$ are 4 and 3, respectively (see the "Mean of $b_j$" row in either panel of Table 3). The

**Table 3** What May Be Learned From a 2-by-2 Factorial Design

(a)

| | | Room Size (B) | | | | |
|---|---|---|---|---|---|---|
| | | *Small ($b_1$)* | *Large ($b_2$)* | *Mean of $a_i$* | *Main Effect of A* | *Simple Effect of B at $a_i$* |
| Room Color (A) | Pink ($a_1$) | (i) Small, Pink ($ab_{11}$) 5 | (ii) Large, Pink ($ab_{12}$) 4 | (5 + 4) ÷ 2 = 4.5 | 4.5 − 2.5 = 2 | At $a_1$: $d_3 = (5-4) = 1$ |
| | White ($a_2$) | (iii) Small, White ($ab_{21}$) 3 | (iv) Large, White ($ab_{22}$) 2 | (3 + 2) ÷ 2 = 2.5 | | At $a_2$: $d_4 = (3-2) = 1$ |
| Mean of $b_j$ | | (5 + 3) ÷ 2 = 4 | (4 + 2) ÷ 2 = 3 | $(D_{of}D)_1$: $d_1 - d_2 = 2 - 2 = 0$ | | |
| | | | | $(D_{of}D)_2$: $d_3 - d_4 = 1 - 1 = 0$ | | |
| Main effect of B | | 4 − 3 = 1 | | [Q1]: Is $(D_{of}D)_{12}$ zero? | | |
| Simple effect of A at $b_j$ | | At $b_1$: $d_1 = (5-3) = 2$ | At $b_2$: $d_2 = (4-2) = 2$ | [Q2]: Is $(D_{of}D)_{34}$ zero? | | |

(b)

| | | Room Size (B) | | | | |
|---|---|---|---|---|---|---|
| | | *Small ($b_1$)* | *Large ($b_2$)* | *Mean of $a_i$* | *Main Effect of A* | *Simple Effect of B at $a_i$* |
| Room Color (A) | Pink ($a_1$) | (i) Small, Pink ($ab_{11}$) 15 | (ii) Large, Pink ($ab_{12}$) 7 | (15 + 7) ÷ 2 = 11 | 11 − 6 = 5 | At $a_1$: $d_3 = (15-7) = 8$ |
| | White ($a_2$) | (iii) Small, White ($ab_{21}$) 2 | (iv) Large, White ($ab_{22}$) 10 | (2 + 10) ÷ 2 = 6 | | At $a_2$: $d_4 = (2-10) = -8$ |
| Mean of $B_j$ | | (15 + 2) ÷ 2 = 8.5 | (7 + 10) ÷ 2 = 8.5 | $(D_{of}D)_1$: $d_1 - d_2 = 13 - (-3) = 16$ | | |
| | | | | $(D_{of}D)_2$: $d_3 - d_4 = 8 - (-8) = 16$ | | |
| Main effect of B | | 8.5 − 8.5 = 0 | | [Q1]: Is $(D_{of}D)_{12}$ zero? | | |
| Simple effect of A at $B_j$ | | At $b_1$: $d_1 = (15-2) = 13$ | At $b_2$: $d_2 = (7-10) = -3$ | [Q2]: Is $(D_{of}D)_{34}$ zero? | | |

*Notes*: (a) An example of additive effects of *A* and *B*. (b) An example of *AB* interaction (nonadditive) effects.

main effect of *B* is 1. That is, the two levels of *B* (or *A*) are averaged when the main effect of *A* (or *B*) is being considered.

### Simple Effect

Given that there are two levels of *A* (or *B*), it is possible to ask whether or not the two levels of *B* (or *A*) differ at either level of *A* (or *B*). Hence, there are the entries, $d_3$ and $d_4$, in the "Simple effect of *B* at $a_i$" column, and the entries, $d_1$ and $d_2$, "Simple effect of *A* at $b_j$" row in either panel of Table 3. Those entries are the four simple effects of the 2-by-2 factorial experiment. They may be summarized as follows:

$d_1$ = Simple effect of *A* at $b_1$ is $(ab_{11} - ab_{21}) = (5 - 3) = 2$;

$d_2$ = Simple effect of *A* at $b_2$ is $(ab_{12} - ab_{22}) = (4 - 2) = 2$;

$d_3$ = Simple effect of *B* at $a_1$ is $(ab_{12} - ab_{11}) = (4 - 5) = -1$;

$d_4$ = Simple effect of *B* at $a_2$ is $(ab_{22} - ab_{21}) = (2 - 3) = -1$.

### AB Interaction

In view of the fact that there are two simple effects of *A* (or *B*), it is important to know whether or not they differ. Consequently, the effects noted above give rise to the following questions:

[Q1] $(D_{of}D)_{12}$: Is $d_1 - d_2 = 0$?

[Q2] $(D_{of}D)_{34}$: Is $d_3 - d_4 = 0$?

Given that $d_1 - d_2 = 0$, one is informed that the effect of Variable *A* is independent of that of Variable *B*. By the same token, that $d_3 - d_4 = 0$ means that the effect of Variable *B* is independent of that of Variable *A*. That is to say, when the answers to both [Q1] and [Q2] are "Yes," the joint effects of Variables *A* and *B* on the dependent variable are the sum of the individual effects of Variables *A* and *B*. Variables *A* and *B* are said to be *additive* in such an event.

Panel (b) of Table 3 illustrates a different scenario. The answers to both [Q1] and [Q2] are "No." It informs one that the effects of Variable *A* (or *B*) on the dependent variable differ at different levels of Variable *B* (or *A*). In short, it is learned from a "No" answer to either [Q1] or [Q2] (or both) that the joint effects of Variables *A* and *B* on the dependent variables are *nonadditive* in the sense that their joint effects are *not* the simple sum of the two separate effects. Variables *A* and *B* are said to interact (or there is a two-way *AB* interaction) in such an event.

### *Multifactor Designs*

What has been said about two-factor designs also applies to designs with three or more independent variables (i.e., multifactor designs). For example, in the case of a three-factor design, it is possible to ask questions about three main effects (*A, B,* and *C*); three 2-way interaction effects (*AB, AC,* and *BC* interactions); a set of *simple effects* (e.g., the effect of Variable *C* at different treatment combinations of *AB,* etc.); and a three-way interaction (viz., *ABC* interaction).

## Genres of Experimental Designs for Data Interpretation Purposes

Experimental designs may also be classified in terms of how subjects are assigned to the treatment combinations, namely, *completely randomized, repeated measures, randomized block,* and *split-plot.*

### *Completely Randomized Design*

Suppose that there are 36 prospective subjects. As it is always advisable to assign an equal number of subjects to each treatment combination, six of them are assigned randomly to each of the six treatment combinations of a 2-by-3 factorial experiment. It is called the *completely randomized* design, but more commonly known as an *unrelated sample* (or an *independent sample*) design when there are only two levels to a lone independent variable.

### *Repeated Measures Design*

All subjects are tested in all treatment combinations in a repeated measures design. It is known by the more familiar name *related samples* or *dependent samples* design when there are only two levels to a lone independent variable. The related samples case may be used to illustrate one complication, namely, the potential artifact of the order of testing effect.

Suppose that all subjects are tested at Level I (or II) before being tested at Level II (or I). Whatever the outcome might be, it is not clear whether the result is due to an inherent difference between Levels I and II or to the proactive effects of the level used first on the performance at the subsequent level of the independent variable. For this reason, a procedure is used to balance the order of testing.

Specifically, subjects are randomly assigned to two subgroups. Group 1 is tested with one order (e.g., Level I before Level II), whereas Group 2 is tested with the other order (Level II before Level I). The more sophisticated Latin square arrangement is used to balance the order of test when there are three or more levels to the independent variable.

### *Randomized Block Design*

The nature of the levels used to represent an independent variable may preclude the use of the repeated measures design. Suppose that the two levels of therapeutic method are surgery and radiation. As either of these levels has irrevocable consequences, subjects cannot be used in both conditions. Pairs of subjects have to be selected, assigned, and tested in the following manner.

Prospective subjects are first screened in terms of a set of relevant variables (body weight, severity of symptoms, etc.). Pairs of subjects who are identical (or similar within acceptable limits) are formed. One member of each pair is assigned randomly to surgery, and the other member to

**Table 4** Inductive Principles Beyond the Method of Difference

(a)

| | | *Control Variables* | | | | *Control Procedure* | |
|---|---|---|---|---|---|---|---|
| *Test Condition* | *Independent Variable Manipulated, Medication* | *Age* | *Health* | *Sex* | *IQ* | *Random Assignment of Subjects* | *Dependent Variable, Affective Score* |
| Experimental (High dose) | 10 units | Middle-aged | Good | Male | Normal | $S_1, S_{21}, S_7, \ldots S_{36}$ | To be collected and analyzed |
| Experimental (Low dose) | 5 units | Middle-aged | Good | Male | Normal | $S_9, S_{10}, S_{24}, \ldots S_{27}$ | |
| Control | Placebo | Middle-aged | Good | Male | Normal | $S_9, S_{10}, S_{24}, \ldots S_{12}$ | |

(b)

| | | *Control Variables* | | | | *Control Procedure* | |
|---|---|---|---|---|---|---|---|
| *Test Condition* | *Independent Variable Manipulated, Wall Color* | *Age* | *Health* | *Sex* | *IQ* | *Random Assignment of Subjects* | *Dependent Variable, Affective Score* |
| Experimental | Pink | Middle-aged | Good | Male | Normal | $S_1, S_{21}, S_7, \ldots S_{15}$ | To be collected and analyzed |
| Control (hue) | White | Middle-aged | Good | Male | Normal | $S_9, S_{10}, S_{24}, \ldots S_2$ | |
| Control (brightness) | Green | Middle-aged | Good | Male | Normal | $S_9, S_{10}, S_{24}, \ldots S_{12}$ | |

*Notes*: (a) Method of concomitant variation. (b) Joint method of agreement and difference.

radiation. This matched-pair procedure is extended to matched triplets (or groups of four subjects matched in terms of a set of criteria) if there are three (or four) levels to the independent variable. Each member of the triplets (or four-member groups) is assigned randomly to one of the treatment combinations.

### *Split-Plot Design*

A *split-plot* design is a combination of the repeated measures design and the completely randomized design. It is used when the levels of one of the independent variables has irrevocable effects (e.g., surgery or radiation of therapeutic method), whereas the other independent variable does not (e.g., Drugs A and B of type of drug).

## Underlying Inductive Logic

Designs other than the one-factor, two-level design implicate two other rules of induction, namely, the *method of concomitant variation* and the *joint method of agreement and difference*.

### *Method of Concomitant Variation*

Consider a study of the effects of a drug's dosage. The independent variable is *dosage,* whose three levels are 10, 5, and 0 units of the medication in question. As *dosage* is a quantitative variable, it is possible to ask whether or not the effect of treatment varies systematically with dosage. The experimental conditions are arranged in the way shown in Panel (a) of Table 4 that depicts the method of concomitant variation.

The control variables and procedures in Tables 1 and 4 are the same. The only difference is that each row in Table 4 represents a level (of a single independent variable) or a treatment combination (when there are two or more independent variables). That is to say, the method of concomitant variation is the logic underlying factorial designs of any size when quantitative independent variables are used.

### *Joint Method of Agreement and Difference*

Shown in Panel (b) of Table 4 is the joint method of agreement and disagreement. Whatever is true of Panel (a) of Table 4 also applies to Panel (b) of

Table 4. It is the underlying inductive rule when a qualitative independent variable is used (e.g., *room color*).

In short, an experimental design is a stipulation of the formal arrangement of the independent, control, and independent variables, as well as the control procedure, of an experiment. Underlying every experimental design is an inductive rule that reduces ambiguity by rendering it possible to exclude alternative interpretations of the result. Each control variable or control procedure excludes one alternative explanation of the data.

*Siu L. Chow*

*See also* Replication; Research Hypothesis; Rosenthal Effect

**Further Readings**

Boring, E. G. (1954). The nature and history of experimental control. *American Journal of Psychology, 67,* 573–589.

Chow, S. L. (1992). *Research methods in psychology: A primer.* Calgary, Alberta, Canada: Detselig.

Mill, J. S. (1973). *A system of logic: Ratiocinative and inductive.* Toronto, Ontario, Canada: University of Toronto Press.

# Experimenter Expectancy Effect

The experimenter's expectancy effect is an important component of the social psychology of the psychological experiment (SPOPE), whose thesis is that conducting or participating in research is a social activity that might be affected subtly by three social or interpersonal factors, namely, demand characteristics, subject effects, and the experimenter's expectancy effects. These artifacts call into question the credibility, generality, and objectivity, respectively, of research data. However, these artifacts may be better known as social psychology of nonexperimental research (SPONE) because they apply only to nonexperimental research.

## The SPOPE Argument

Willing to participate and being impressed by the aura of scientific investigation, research participants may do whatever is required of them. This *demand characteristics* artifact creates credibility issues in the research data. The *subject effect* artifact questions the generalizability of research data. This issue arises because participants in the majority of psychological research are volunteering tertiary-level students who may differ from the population at large.

As an individual, a researcher has profound effects on the data. Any personal characteristics of the researcher may affect research participants (e.g., ethnicity, appearance, demeanor). Having vested interests in certain outcomes, researchers approach their work from particular theoretical perspectives. These biases determine in some subtle and insidious ways how researchers might behave in the course of conducting research. This is the *experimenter expectancy effect* artifact.

At the same time, the demand characteristics artifact predisposes research participants to pick up cues about the researcher's expectations. Being obligingly ingratiatory, research participants "cooperate" with the researcher to obtain the desired results. The *experimenter expectancy effect* artifact detracts research conclusions from their objectivity.

## SPOPE Revisited—SPONE

### *Limits of Goodwill*

Although research participants bear goodwill toward researchers, they may not (and often cannot) fake responses to please the researcher as implied in the SPOPE thesis.

To begin with, research participants might give untruthful responses only when illegitimate features in the research procedure render it necessary and possible. Second, it is not easy to fake responses without being detected by the researcher, especially when measured with a well-defined task (e.g., the attention span task). Third, it is not possible to fake performance that exceeds the participants' capability.

### *Nonexperiment Versus Experiment*

Faking on the part of research participants is not an issue when experimental conclusions are based on subjects' differential performance on the attention span task in two or more conditions with proper controls. Suppose that a properly selected

**Table 1** The Basic Structure of an Experiment

| | | Control Variable | | | | |
|---|---|---|---|---|---|---|
| *Test Condition* | *Independent Variable* | *IQ* | *Sex* | *Age* | *Control Procedure* | *Dependent Variable* |
| Experimental | Drug | Normal | M | 12–15 | Random assignment | Longer attention span |
| Control | Placebo | Normal | M | 12–15 | Repeated measures | Shorter attention span |

**Table 2** A Schematic Representation of the Design of Goldstein, Rosnow, Goodstadt, and Sul's (1972) Study of Verbal Conditioning

| *Subject Group* | *Experimental Condition (Knowledgeable of verbal conditioning)* | *Control Condition (Not knowledgeable of verbal conditioning)* | *Mean of Two Means* | *Difference Between Two Means* |
|---|---|---|---|---|
| Volunteers | $\overline{X}_1$ (6) | $\overline{X}_2$ (3) | $\overline{X}$ (4.5) | $d_1 = \overline{X}_1 - \overline{X}_2(6-3) = 3$ |
| Nonvolunteers | $\overline{Y}_1$ (4) | $\overline{Y}_2$ (1) | $\overline{Y}$ (2.5) | $d_2 = \overline{Y}_1 - \overline{Y}_2(4-1) = 3$ |

*Source:* Goldstein, J. J., Rosnow, R. L., Goodstadt, B., & Suls, J. M. (1972). The "good subject" in verbal operant conditioning research. *Journal of Experimental Research in Personality, 6,* 29–33.

*Notes:* Hypothetical mean increase in the number of first-person pronouns used. The numbers in parentheses are added for illustrative purposes only. They were not Goldstein et al.'s (1972) data.

sample of boys is assigned randomly to the two conditions in Table 1. Further suppose that one group fakes to do well, and the other fakes to do poorly. Nonetheless, it is unlikely that the said unprincipled behavior would produce the difference between the two conditions desired by the experimenter.

### *Difference Is Not Absolute*

Data obtained from college or university students do not necessarily lack generality. For example, students also have two eyes, two ears, one mouth, and four limbs like typical humans have. That is, it is not meaningful to say simply that *A* differs from *B*. It is necessary to make explicit (a) the dimension on which *A* and *B* differ, and (b) the relevancy of the said difference to the research in question.

It is also incorrect to say that researchers employ tertiary students as research participants simply because it is convenient to do so. On the contrary, researchers select participants from special populations in a theoretically guided way when required. For example, they select boys with normal IQ within a certain age range when they study hyperactivity. More important, experimenters assign subjects to test conditions in a theoretically guided way (e.g., completely random).

### *Individual Differences Versus Their Effects on Data*

Data shown in Table 2 have been used to support the subject effect artifact. The experiment was carried out to test the effect of volunteering on how fast one could learn. Subjects were verbally reinforced for uttering first-person pronouns. Two subject variables are used (*volunteering status* and *knowledgeability of conditioning principle*).

Of interest is the statistically significant main effect of *volunteering status*. Those who volunteered were conditioned faster than those who did not. Note that the two levels of any subject variable are, by definition, different. Hence, the significant main effect of *volunteering status* is not surprising (see the "Mean of Two Means" column in Table 2). It merely confirms a pre-existing individual difference, but not the required effect of individual differences on experimental data. The data do not support the subject effect artifact because the required two-way interaction between *volunteering status* and *knowledgeability of conditioning* is not significant.

### *Experiment Versus Meta-Experiment*

R. Rosenthal and K. L. Fode were the investigators in Table 3 who instructed *A, B, C,* and *D*

**Table 3** The Design of Rosenthal and Fode's (1963) Experimental Study of Expectancy

| | *Investigators (Rosenthal & Fode)* | | | | | |
|---|---|---|---|---|---|---|
| *Expectation Group* | *+5 Data Collectors* | | | *−5 Data Collectors* | | |
| Data collector | *A* | ... | *B* | *C* | ... | *D* |
| Subject | $i_1$ | ... | $j_1$ | $k_1$ | ... | $q_1$ |
| | $i_2$ | ... | $j_2$ | $k_2$ | ... | $q_2$ |
| | ... | ... | ... | ... | ... | ... |
| | ... | ... | ... | ... | ... | ... |
| | $i_n$ | ... | $j_n$ | $k_n$ | ... | $q_n$ |
| Mean rating collected by individual data collector | [*x*] | ... | [*y*] | [*z*] | ... | [*w*] |
| Mean rating of data collectors as a group | | 4.05 | | | −0.95 | |

*Source:* Rosenthal, R., & Fode, K. L. (1963). Three experiments in experimenter bias. *Psychological Reports, 12,* 491–511.

**Table 4** The Design of Chow's (1994) Meta-Experiment

(a)

| | *Investigator (Chow, 1994)* | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Expectancy Group* | *+5 Expectancy* | | | | *No Expectancy* | | | | *−5 Expectancy* | | | |
| *Experimenter* | *A* | | *B* | | *F* | | *G* | | *P* | | *Q* | |
| Test condition (H = Happy face) (S = Sad face) | H | S | H | S | H | S | H | S | H | S | H | S |
| Subjects | $i_1$ | $i'_1$ | $j_1$ | $j'_1$ | $k_1$ | $k'_1$ | $m_1$ | $m'_1$ | $n_1$ | $n'_1$ | $q_1$ | $q'_1$ |
| | $i_2$ | $i'_2$ | $j_2$ | $j'_2$ | $k_2$ | $k'_2$ | $m_2$ | $m'_2$ | $n_2$ | $n'_2$ | $q_2$ | $q'_2$ |
| | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| | $i_n$ | $i'_n$ | $j_n$ | $j'_n$ | $k_n$ | $k'_n$ | $m_n$ | $m'_n$ | $n_n$ | $n'_n$ | $q_n$ | $q'_n$ |
| Mean rating collected by individual data collector | [*x*] | [*x'*] | [*y*] | [*y'*] | [*w*] | [*w'*] | [*z*] | [*w'*] | [*a*] | [*a'*] | [*b*] | [*b'*] |

(b)

| *Expectancy Group* | *+5 Expectancy* | | *No Expectancy* | | *−5 Expectancy* | |
|---|---|---|---|---|---|---|
| Stimulus face | Happy | Sad | Happy | Sad | Happy | Sad |
| Mean | 3.16 | −0.89 | 3.88 | −0.33 | 2.95 | −0.38 |
| Difference between two means: $(\overline{E} - \overline{C})$ | 4.05 | | 4.21 | | 3.33 | |

*Source:* Chow, S. L. (1994). The experimenter's expectancy effect: A meta experiment. *Zeitschrift für Pädagogische Psychologie* (German Journal of Educational Psychology), *8,* 89–97.

*Notes:* (a) The design of the meta-experiment. (b) Mean ratings in the experimental ("Happy") and control ("Sad") condition as a function of expectancy.

to administer a photo-rating task to their own groups of participants. Participants had to indicate whether or not a photograph (with a neutral expression) was one of a successful person or an unsuccessful person. The investigator induced *A* and *B* (see the "+5 Data Collectors" columns) to expect a mean rating of +5 ("successful") from their participants, and *C* and *D* were induced to expect a mean rating of −5 ("unsuccessful"; see the "−5 Data Collectors" column). The observed difference between the two groups of data collectors is deemed consistent with the

experimenter expectancy effect artifact (i.e., 4.05 versus −0.95).

Although Rosenthal and Fode are experimenters, *A, B, C,* and *D* are not. All of them collected absolute measurement data in one condition only, not collecting experimental data of differential performance. To test the experimenter expectancy effect artifact in a valid manner, the investigators in Table 3 must give each of *A, B, C,* and *D* an experiment to conduct. That is, the experimenter expectancy effect artifact must be tested with a meta-experiment (i.e., an experiment about experiment), an example of which is shown in panel (a) of Table 4.

Regardless of the expectancy manipulation (positive [+5], neutral [0], or negative [−5]), Chow gave each of *A, B, F, G, P,* and *Q* an experiment to conduct. That is, every one of them obtained from his or her own group of subjects the differential performance on the photo-rating task between two conditions (Happy Face vs. Sad Face).

It is said in the experimenter expectancy effect argument that subjects behave in the way the experimenter expects. As such, that statement is too vague to be testable. Suppose that a sad face was presented. Would both the experimenter (e.g., *A* or *Q* in Table 4) and subjects (individuals tested by *A* or *Q*) ignore that it was a sad (or happy) face and identify it as "successful" (or "unsuccessful") under the "+5" (or "−5") condition? Much depends on the consistency between *A*'s or *Q*'s expectation and the nature of the stimulus (e.g., happy or sad faces), as both *A* (or *Q*) and his or her subjects might moderate or exaggerate their responses.

### Final Thoughts

SPOPE is so called because the distinction between experimental and nonexperimental empirical research has not been made as a result of not appreciating the role of control in empirical research. Empirical research is an experiment only when three control features are properly instituted (a valid comparison baseline, constancy of conditions, and procedures for eliminating artifacts). As demand characteristics, participant effect and expectancy effect may be true of nonexperimental research in which no control is used, in which case it is more appropriate to characterize the SPOPE phenomenon as SPONE. Those putative artifacts are not applicable to experimental studies.

*Siu L. Chow*

***See also*** Experimental Design

**Further Readings**

Chow, S. L. (1992). *Research methods in psychology: A primer.* Calgary, Alberta, Canada: Detselig.

Chow, S. L. (1994). The experimenter's expectancy effect: A meta experiment. *Zeitschrift für Pädagogische Psychologie* (German Journal of Educational Psychology), *8,* 89–97.

Goldstein, J. J., Rosnow, R. L., Goodstadt, B., & Suls, J. M. (1972). The "good subject" in verbal operant conditioning research. *Journal of Experimental Research in Personality, 6,* 29–33.

Orne, M. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist, 17,* 776–783.

Rosenthal, R., & Fode, K. L. (1963). Three experiments in experimenter bias. *Psychological Reports, 12,* 491–511.

## Exploratory Data Analysis

Exploratory data analysis (EDA) is a data-driven conceptual framework for analysis that is based primarily on the philosophical and methodological work of John Tukey and colleagues, which dates back to the early 1960s. Tukey developed EDA in response to psychology's overemphasis on hypodeductive approaches to gaining insight into phenomena, whereby researchers focused almost exclusively on the hypothesis-driven techniques of confirmatory data analysis (CDA). EDA was not developed as a substitute for CDA; rather, its application is intended to satisfy a different stage of the research process. EDA is a bottom-up approach that focuses on the initial exploration of data; a broad range of methods are used to develop a deeper understanding of the data, generate new hypotheses, and identify patterns in the data. In contrast, CDA techniques are of greater value at a later stage when the emphasis is on

testing previously generated hypotheses and confirming predicted patterns. Thus, EDA offers a different approach to analysis that can generate valuable information and provide ideas for further investigation.

## Ethos

A core goal of EDA is to develop a detailed understanding of the data and to consider the processes that might produce such data. Tukey used the analogy of EDA as detective work because the process involves the examination of facts (data) for clues, the identification of patterns, the generation of hypotheses, and the assessment of how well tentative theories and hypotheses fit the data.

EDA is characterized by flexibility, skepticism, and openness. Flexibility is encouraged as it is seldom clear which methods will best achieve the goals of the analyst. EDA encourages the use of statistical and graphical techniques to understand data, and researchers should remain open to unanticipated patterns. However, as summary measures can conceal or misrepresent patterns in data, EDA is also characterized by skepticism. Analysts must be aware that different methods emphasize some aspects of the data at the expense of others; thus, the analyst must also remain open to alternative models of relationships.

If an unexpected data pattern is uncovered, the analyst can suggest plausible explanations that are further investigated using confirmatory techniques. EDA and CDA can supplement each other: Where the abductive approach of EDA is flexible and open, allowing the data to drive subsequent hypotheses, the more ambitious and focused approach of CDA is hypothesis-driven and facilitates probabilistic assessments of predicted patterns. Thus, a balance is required between an exploratory and confirmatory lens being applied to data; EDA comes first, and ideally, any given study should combine both.

## Methods

EDA techniques are often classified in terms of the four *R*s: revelation, residuals, reexpression, and resistance. However, it is not the use of a technique per se that determines whether it is EDA, but the *purpose* for which it is used—namely, to assist the development of rich mental models of the data.

### *Revelation*

EDA encourages the examination of different ways of describing the data to understand inherent patterns and to avoid being fooled by unwarranted assumptions.

#### Data Description

The use of summary descriptive statistics offers a concise representation of data. EDA relies on *resistant statistics,* which are less affected by deviant cases. However, such statistics involve a tradeoff between being concise versus precise; therefore, an analyst should never rely exclusively on statistical summaries. EDA encourages analysts to examine data for skewness, outliers, gaps, and multiple peaks, as these can present problems for numerical measures of spread and location. Visual representations of data are required to identify such instances to inform subsequent analyses. For example, based on their relationship to the rest of the data, outliers may be omitted or may become the focus of the analysis, a distribution with multiple peaks may be split into different distributions, and skewed data may be *reexpressed.* Inadequate exploration of the data distribution through visual representations can result in the use of descriptive statistics that are not characteristic of the entire set of values.

#### Data Visualization

Visual representations are encouraged because graphs provide parsimonious representations of data that facilitate the development of suitable mental models. Graphs display information in a way that makes it easier to detect unexpected patterns. EDA emphasizes the importance of using numerous graphical methods to see what each reveals about the data structure.

Tukey developed a number of EDA graphical tools, including the *box-and-whisker plot,* otherwise known as the *box plot.* Box plots are useful for examining data and identifying potential outliers; however, like all data summarization methods, they focus on particular aspects of the data. Therefore, other graphical methods should also be used. *Stem-and-leaf displays* provide valuable additional

information because all data are retained in a frequency table, providing a sense of the distribution shape. In addition, *dot plots* highlight gaps or dense parts of a distribution and can identify outliers.

Tukey's emphasis on graphical data analysis has influenced statistical software programs, which now include a vast array of graphical techniques. These techniques can highlight individual values and their relative position to each other, check data distributions, and examine relationships between variables and relationships between regression lines and actual data. In addition, interactive graphics, such as linked plots, allow the researcher to select a specific case or cases in one graphical display (e.g., scatterplot) and see the same case(s) in another display (e.g., histogram). Such an approach could identify cases that are bivariate outliers but not outliers on either of the two variables being correlated.

### *Residuals*

According to Tukey, the idea of data analysis is explained using the following formula: DATA = SMOOTH + ROUGH, or, more formally, DATA = FIT + RESIDUALS, based on the idea that the way in which we describe/model data is never completely accurate because there is always some discrepancy between the model and the actual data. The smooth is the underlying, simplified pattern in the data; for example, a straight line representing the relationship between two variables. However, as data never conform perfectly to the smooth, deviations from the smooth (the model) are termed the rough (the residuals).

Routine examination of residuals is one of the most influential legacies of EDA. Different models produce different patterns of residuals; consequently, examining residuals facilitates judgment of a model's adequacy and provides the means to develop better models. From an EDA perspective, the rough is just as important as the smooth and should never be ignored.

Although *residual sums-of-squares* are widely used as a measure of the discrepancy between the model and the data, relying exclusively on this measure of model fit is dangerous as important patterns in the residuals may be overlooked. Detailed examination of the residuals reveals valuable information about a model's misspecifications. EDA thus emphasizes careful examination of residual plots for any additional patterns, such as curves or multiple modes, as this suggests that the selected model failed to describe an important aspect of the data. In such instances, further smoothing is required to get at the underlying pattern.

EDA focuses on building models and generating hypotheses in an iterative process of model specification. The analyst must be open to alternative models, and thus the residuals of different models are examined to see if there is a better fit to the data. Thus, models are generated, tested, modified, and retested in a cyclical process that should lead the researcher, by successive approximation, toward a good description of the data. Model building and testing require heeding data at all stages of research, especially the early stages of analysis. After understanding the structure of each variable separately, pairs of variables are examined in terms of their patterns, and finally, multivariate models of data can be built iteratively. This iterative process is integral to EDA's ethos of using the data to develop and refine models.

Suitable models that describe the data adequately can then be compared to models specified by theory. Alternatively, EDA can be conducted on one subset of data to generate models, and then confirmatory techniques can be applied subsequently to test these models in another subset. Such cross-validation means that when patterns are discovered, they are considered provisional until their presence is confirmed in a different data set.

### *Reexpression*

Real-life data are often messy, and EDA recognizes the importance of scaling data in an appropriate way so that the phenomena are represented in a meaningful manner. Such data transformation is referred to as data *reexpression* and can reveal additional patterns in the data. Reexpression can affect the actual numbers, the relative distances between the values, and the rank ordering of the numbers. Thus, EDA treats measurement scales as arbitrary, advocating a flexible approach to examination of data patterns.

Reexpression may make data suitable for parametric analysis. For example, nonlinear transformations (e.g., log transformation) can make data

follow a normal distribution or can stabilize the variances. Reexpression can result in linear relationships between variables that previously had a nonlinear relationship. Making the distribution symmetrical about a single peak makes modeling of the data pattern easier.

### *Resistance*

*Resistance* involves the use of methods that minimize the influence of extreme or unusual data. Different procedures may be used to increase resistance. For example, absolute numbers or rank-based summary statistics can be used to summarize information about the shape, location, and spread of a distribution instead of measures based on sums. A common example of this is the use of the median instead of the mean; however, other resistant central tendency measures can also be used, such as the *trimean* (the mean of the 25th percentile, the 75th percentile, and the median counted twice). Resistant measures of spread include the interquartile range and the median absolute deviation.

Resistance is increased by giving greater weight to values that are closer to the center of the distribution. For example, a *trimmed mean* may be used whereby data points above a specified value are excluded from the estimation of the mean. Alternatively, a *Winsorized mean* may be used where the tail values of a distribution are pulled in to match those of a specified extreme score.

### Outliers

Outliers present the researcher with a choice between including these extreme scores (which may result in a poor model of all the data) and excluding them (which may result in a good model that applies only to a specific subset of the original data). EDA considers why such extreme values arise. If there is evidence that outliers were produced by a different process from the one underlying the other data points, it is reasonable to exclude the outliers as they do not reflect the phenomena under investigation. In such instances, the researcher may need to develop different models to account for the outliers. For example, outliers may reflect different subpopulations within the data set.

However, often there is no clear reason for the presence of outliers. In such instances, the impact of outliers is examined by comparing the residuals from a model based on the entire data set with one that excludes outliers. If results are consistent across the two models, then either course of action may be followed. Conversely, if substantial differences exist, then both models should be reported and the impact of the outliers needs to be considered. From an EDA perspective, the question is not how to deal with outliers but what can be learned from them. Outliers can draw attention to important aspects of the data that were not originally considered, such as unanticipated psychological processes, and provide feedback regarding model misspecification. This data-driven approach to improving models is inherent in the iterative EDA process.

## Conclusion

EDA is a data-driven approach to gaining familiarity with data. It is a distinct way of thinking about data analysis, characterized by an attitude of flexibility, openness, skepticism, and creativity to discovering patterns, avoiding errors, and developing useful models that are closely aligned to data. Combining insights from EDA with the powerful analytical tools of CDA provides a robust approach to data analysis.

*Maria M. Pertl and David Hevey*

***See also*** Box-and-Whisker Plot; Histogram; Outlier; Residual Plot; Residuals; Scatterplot

### Further Readings

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (Eds.). (1983). *Understanding robust and exploratory data analysis*. New York: John Wiley.

Keel, T. G., Jarvis, J. P., & Muirhead, Y. E. (2009). An exploratory analysis of factors affecting homicide investigations: Examining the dynamics of murder clearance rates. *Homicide Studies, 13,* 50–68.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

# Exploratory Factor Analysis

Exploratory factor analysis (EFA) is a multivariate statistical technique to model the covariance

structure of the observed variables by three sets of parameters: (a) factor loadings associated with latent (i.e., unobserved) variables called factors, (b) residual variances called unique variances, and (c) factor correlations. EFA aims at explaining the relationship of many observed variables by a relatively small number of factors. Thus, EFA is considered one of the data reduction techniques. Historically, EFA dates back to Charles Spearman's work in 1904, and the theory behind EFA has been developed along with the psychological theories of intelligence, such as L. L. Thurstone's multiple factor model. Today, EFA is among the most frequently used statistical techniques by researchers in the social sciences and education.

It is well-known that EFA often gives the solution similar to principal component analysis (PCA). However, there is a fundamental difference between EFA and PCA in that factors are predictors in EFA, whereas in PCA, principal components are outcome variables created as a linear combination of observed variables. Here, an important note is that PCA is a different method from principal factor analysis (also called the principal axis method). Statistical software such as IBM® SPSS® (PASW) 18.0 (an IBM company, formerly named PASW® Statistics) supports both PCA and principal factor analysis. Another similarity exists between EFA and confirmatory factor analysis (CFA). In fact, CFA was developed as a variant of EFA. The major difference between EFA and CFA is that EFA is typically employed without prior hypotheses regarding the covariance structure, whereas CFA is employed to test the prior hypotheses on the covariance structure. Often, researchers do EFA and then do CFA using a different sample. Note that CFA is a submodel of structural equation models. It is known that two-parameter item response theory (IRT) is mathematically equivalent to the one-factor EFA with ordered categorical variables. EFA with binary and ordered categorical variables can also be treated as a generalized latent variable model (i.e., a generalized linear model with latent predictors).

Mathematically, EFA expresses each observed variable ($x_i$) as a linear combination of factors ($f_1$, $f_2, \ldots, f_m$) plus an error term, that is, $x_i = \mu_i + \lambda_{i1}f_1 + \lambda_{i2}f_2 + \cdots + \lambda_{im}f_m + e_i$, where $m$ is the number of factors, $\mu_i$ is the population mean of $x_i$,$\lambda_{ij}$s are called the factor loadings or factor patterns, and $e_i$ contains measurement errors and uniqueness. It is almost like a multiple regression model; however, the major difference from multiple regression is that in EFA, the factors are latent variables and not observed. The model for EFA is often given in a matrix form:

$$\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{\Lambda f} + \boldsymbol{e}, \tag{1}$$

where $\boldsymbol{x}$, $\boldsymbol{\mu}$, and $\boldsymbol{e}$ are $p$-dimensional vectors, $\boldsymbol{f}$ is an $m$-dimensional vector of factors, and $\boldsymbol{\Lambda}$ is a $p \times m$ matrix of factor loadings. It is usually assumed that factors ($\boldsymbol{f}$) and errors ($\boldsymbol{e}$) are uncorrelated, and different error terms ($e_i$ and $e_j$ for $i \neq j$) are uncorrelated. From the matrix form of the model in Equation 1, we can express the population variance-covariance matrix (covariance structure) $\boldsymbol{\Sigma}$ as

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda\Phi\Lambda}' + \boldsymbol{\Psi} \tag{2}$$

if factors are correlated, where $\boldsymbol{\Phi}$ is an $m \times m$ correlation matrix among factors (factor correlation matrix), $\boldsymbol{\Lambda}'$ is the transpose of matrix $\boldsymbol{\Lambda}$ in which rows and columns of $\boldsymbol{\Lambda}$ are interchanged (so that $\boldsymbol{\Lambda}'$ is a $m \times p$ matrix), and $\boldsymbol{\Psi}$ is a $p \times p$ diagonal matrix (all off-diagonal elements are zero due to uncorrelated $e_i$) of error or unique variances. When the factors ($\boldsymbol{f}$) are not correlated, the factor correlation matrix is equal to the identity matrix (i.e., $\boldsymbol{\Phi} = \boldsymbol{I}_m$) and the covariance structure is reduced to

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda\Lambda}' + \boldsymbol{\Psi} \tag{3}$$

For each observed variable, when factors are not correlated, we can compute the sum of squared factor loadings

$$h_i = \lambda_{i1}^2 + \lambda_{i2}^2 + \cdots + \lambda_{im}^2 = \sum_{j=1}^{m} \lambda_{ij}^2, \tag{4}$$

which is called the communality of the ($i$th) variable. When factors are correlated, the communality is calculated as

$$h_i = \sum_{j=1}^{m} \lambda_{ij}^2 + \sum_{j \neq k} \lambda_{ij}\lambda_{ik}\phi_{jk}. \tag{5}$$

When the observed variables are standardized, the $i$th communality gives the proportion of variability of the $i$th variable explained by the $m$

factors. It is well-known that the squared multiple correlation of the $i$th variable on the remaining $p - 1$ variables gives a lower bound for the $i$th communality.

## Estimation (Extraction)

There are three major estimation methods routinely used in EFA. Each estimation method tries to minimize a distance between the sample covariance matrix $S$ and model-based covariance matrix (estimate of $\boldsymbol{\Sigma}$ based on the EFA model: $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$ because for ease of estimation, initially, the factors are typically assumed to be uncorrelated). The first method tries to minimize the trace (i.e., sum of diagonal elements) of $(1/2)(S - \boldsymbol{\Sigma})^2$ and is called either least-squares (LS) or unweighted least-squares (ULS) method. Although LS is frequently used in multiple regression, it is not so common as an estimation method for EFA because it is not scale invariant. That is, the solution is different if we use the sample correlation matrix or the sample variance-covariance matrix. Consequently, the following two methods (both of which are scale invariant) are frequently used for parameter estimation in EFA. One of them tries to minimize the trace of $(1/2)\{(S - \boldsymbol{\Sigma})S^{-1}\}^2$ and is called the generalized least-squares (GLS) method. Note that $S^{-1}$ is the inverse (i.e., matrix version of reciprocal) of $S$ and serves as a weight matrix here. Another scale-invariant estimation method tries to minimize trace $(S\boldsymbol{\Sigma}^{-1}) - \log(\det(S\boldsymbol{\Sigma}^{-1})) - p$, where det is the determinant operator and log is the natural logarithm. This method is called the maximum-likelihood (ML) method. It is known that when the model holds, GLS and ML give asymptotically (i.e., when sample size is very large) equivalent solutions. In fact, the criterion for ML can be approximated by the trace of $(1/2)[(S - \boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}]^2$, with almost the same function to be minimized as GLS, the only difference being the weight matrix $S^{-1}$ replaced by $\boldsymbol{\Sigma}^{-1}$. When the sample is normally distributed, the ML estimates are asymptotically most efficient (i.e., when the sample size is large, the ML procedure leads to estimates with the smallest variances). Note that the principal factor method frequently employed as an estimation method for EFA is equivalent to ULS when the solution converges. It obtains factor loading estimates using the eigenvalues and eigenvectors of the matrix $\boldsymbol{R} - \boldsymbol{\Psi}$, where $\boldsymbol{R}$ is the sample correlation matrix.

When the factors are uncorrelated, with a $m \times m$ orthogonal matrix $\boldsymbol{T}$ (i.e., $\boldsymbol{T}\boldsymbol{T}' = \boldsymbol{I}_m$), the variance-covariance matrix $\boldsymbol{\Sigma}$ of the observed variables $x$ under EFA given as $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$ can be rewritten as

$$\begin{aligned} \boldsymbol{\Sigma} &= \boldsymbol{\Lambda}\boldsymbol{T}\boldsymbol{T}'\boldsymbol{\Lambda}' + \boldsymbol{\Psi} = (\boldsymbol{\Lambda}\boldsymbol{T})(\boldsymbol{\Lambda}\boldsymbol{T})' + \boldsymbol{\Psi}; \\ \boldsymbol{\Psi} &= \boldsymbol{\Lambda}^*\boldsymbol{\Lambda}^{*\prime} + \boldsymbol{\Psi}, \end{aligned} \tag{6}$$

where $\boldsymbol{\Lambda}^* = \boldsymbol{\Lambda}\boldsymbol{T}$. This indicates that the EFA model has an identification problem called the indeterminacy. That means that we need to impose at least $m(m-1)/2$ constraints on the factor loading matrix in order to estimate the parameters $\lambda_{ij}$ uniquely. For example, in the ML estimation, commonly used constraints are to let $\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}$ be a diagonal matrix. Rotations (to be discussed) are other ways to impose constraints on the factor loading matrix. One can also fix $m(m-1)/2$ loadings in the upper triangle of $\boldsymbol{\Lambda}$ at zero for identification.

In estimation, we sometimes encounter a problem called the improper solution. The most frequently encountered improper solution associated with EFA is that certain estimates of unique variances in $\boldsymbol{\Psi}$ are negative. Such a phenomenon is called the Heywood case. If the improper solution occurs as a result of sampling fluctuations, it is not of much concern. However, it may be a manifestation of model misspecification.

When data are not normally distributed or contain outliers, better parameter estimates can be obtained when the sample covariance matrix $S$ in any of the above estimation methods is replaced by a robust covariance matrix. When a sample contains missing values, the $S$ should be replaced by the maximum-likelihood estimate of the population covariance matrix.

## Number of Factors

We need to determine the number of factors $m$ such that the variance-covariance matrix of observed variables is well approximated by the factor model, and also $m$ should be as small as possible. Several methods are commonly employed to determine the number of factors.

### *Kaiser–Guttman Criterion*

One widely used method is to let the number of factors equal the number of eigenvalues of the sample correlation matrix that are greater than 1. It is called the "eigenvalue-greater-than-1 rule" or the *Kaiser–Guttman criterion*. The rationale behind it is as follows: For a standardized variable, the variance is 1. Thus, by choosing the number of factors equal to eigenvalues that are greater than 1, we can choose the factors whose variance is at least greater than the variance of each (standardized) observed variable. It makes sense from the point of view of data reduction. However, this criterion tends to find too many factors. A variant of the eigenvalue-greater-than-1 rule is the "eigenvalue-greater-than-zero" rule that applies when the diagonal elements of the sample correlation matrix are replaced by the squared multiple correlations when regressing each variable on the rest of the $p - 1$ variables (called the *reduced correlation matrix*).

### *Scree Plot*

Another frequently used rule is a visual plot, with the ordered eigenvalues (from large to small) of the sample correlation matrix in the vertical axis and the ordinal number in the horizontal axis. This plot is commonly called the *scree plot*. It frequently happens with practical data that, after the first few, the eigenvalues taper off as almost a straight line. The number of factors suggested by the scree plot is the number of eigenvalues just before they taper off in a linear fashion. There are variant methods to the scree plot, such as Horn's parallel analysis.

### *Communality*

Because communalities are analogous to the squared multiple correlation in regression, they can be used to aid our decision for the number of factors. Namely, we should choose $m$ such that every communality is sufficiently large.

### *Likelihood Ratio Test*

When the observed sample is normally distributed, the likelihood ratio test (LRT) statistic from the ML procedure can be used to test whether the $m$ factor model is statistically adequate in explaining the relationship of the measured variables. Under the null hypothesis, the LRT statistic asymptotically follows a chi-square distribution with degrees of freedom $df = [(p-m)^2 - (p+m)]/2$. Statistical software reports the LRT statistic with Bartlett correction in which the sample size $n$ is replaced by $n - (2p+5)/6 - 2m/3$, which is supposed to improve the closeness of the LRT to the asymptotic chi-square distribution. A rescaled version of the LRT also exists when data do not follow normal distribution or samples contain missing values.

## Rotation

In his 1947 book, Thurstone argued that the initial solution of factor loadings should be rotated to find a simple structure, that is, the pattern of factor loadings having an easy interpretation. A simple structure can be achieved when, for each row of the factor loading matrix, there is only one element whose absolute value (ignoring the sign) is high and the rest of the elements are close to zero. Several methods for rotation to a simple structure have been proposed. Mathematically, these methods can be regarded as different ways of imposing constraints to resolve the identification problem discussed above.

The rotational methods are classified as orthogonal and oblique rotations. The orthogonal rotations are the rotations in which the factors are uncorrelated, that is, $\boldsymbol{T}\boldsymbol{T}' = \boldsymbol{I}_m$ with the rotation matrix $\boldsymbol{T}$ that connects the initial factor loading matrix $\boldsymbol{A}$ and the rotated factor loading matrix $\boldsymbol{\Lambda}$ such that $\boldsymbol{\Lambda} = \boldsymbol{A}\boldsymbol{T}$. The oblique rotations are the rotations in which the factors are allowed to be correlated with each other. Note here that once we employ an oblique rotation, we need to distinguish between the factor pattern matrix and the factor structure matrix. The factor pattern matrix is a matrix whose elements are standardized regression coefficients of each observed variable on factors, whereas the factor structure matrix represents correlations between the factors and the observed variables. As long as the rotation is orthogonal, the factor pattern and the factor structure matrices are identical, and we often call the identical matrix the factor loading matrix.

**Table 1** Sample Correlation Matrix for the Nine Psychological Tests ($n = 145$)

| | | *x1* | *x2* | *x3* | *x4* | *x5* | *x6* | *x7* | *x8* | *x9* |
|---|---|---|---|---|---|---|---|---|---|---|
| x1 | Visual | **1.000** | | | | | | | | |
| x2 | Cubes | **0.318** | **1.000** | | | | | | | |
| x3 | Flags | **0.468** | **0.230** | **1.000** | | | | | | |
| x4 | Paragraph | **0.335** | **0.234** | **0.327** | **1.000** | | | | | |
| x5 | Sentence | **0.304** | **0.157** | **0.335** | **0.722** | **1.000** | | | | |
| x6 | Word | **0.326** | **0.195** | **0.325** | **0.714** | **0.685** | **1.000** | | | |
| x7 | Addition | **0.116** | **0.057** | **0.099** | **0.203** | **0.246** | **0.170** | **1.000** | | |
| x8 | Counting | **0.314** | **0.145** | **0.160** | **0.095** | **0.181** | **0.113** | **0.585** | **1.000** | |
| x9 | Straight | **0.489** | **0.239** | **0.327** | **0.309** | **0.345** | **0.280** | **0.408** | **0.512** | **1.000** |

**Table 2** SAS Code for EFA for the Data in Table 1

```
data one(type = corr);
input _type_ $ _name_ $ x1-x9;
datalines;
n     .    145   145   145   145   145   145   145   145   145
corr  x1   1.000 .     .     .     .     .     .     .     .
corr  x2   0.318 1.000 .     .     .     .     .     .     .
corr  x3   0.468 0.230 1.000 .     .     .     .     .     .
corr  x4   0.335 0.234 0.327 1.000 .     .     .     .     .
corr  x5   0.304 0.157 0.335 0.722 1.000 .     .     .     .
corr  x6   0.326 0.195 0.325 0.714 0.685 1.000 .     .     .
corr  x7   0.116 0.057 0.099 0.203 0.246 0.170 1.000 .     .
corr  x8   0.314 0.145 0.160 0.095 0.181 0.113 0.585 1.000 .
corr  x9   0.489 0.239 0.327 0.309 0.345 0.280 0.408 0.512 1.000
;
title 'promax solution for nine psychological tests (n = 145)';
proc factor data = one scree method = ml nfactor = 3 rotate = promax se;
run;
```

**Table 3** Eigenvalues of the Sample Correlation Matrix

| | *Eigenvalue* | *Difference* | *Proportion* | *Cumulative Proportion* |
|---|---|---|---|---|
| 1 | **3.557** | 1.977 | 0.395 | 0.395 |
| 2 | **1.579** | 0.423 | 0.176 | 0.571 |
| 3 | **1.156** | 0.367 | 0.128 | 0.699 |
| 4 | 0.789 | 0.218 | 0.088 | 0.787 |
| 5 | 0.571 | 0.142 | 0.063 | 0.850 |
| 6 | 0.429 | 0.069 | 0.048 | 0.898 |
| 7 | 0.360 | 0.054 | 0.040 | 0.938 |
| 8 | 0.306 | 0.052 | 0.034 | 0.972 |
| 9 | 0.254 | | 0.028 | 1.000 |

*Note:* The squared multiple correlations are $x_1$: 0.408, $x_2$: 0.135, $x_3$: 0.274, $x_4$: 0.631, $x_5$: 0.600, $x_6$: 0.576, $x_7$: 0.402, $x_8$: 0.466, $x_9$: 0.437.

Among the orthogonal rotations, Kaiser's varimax rotation is by far the most frequently used. The varimax rotation tries to achieve a simple structure by rotating to maximize the variance of squared factor loadings. More specifically, letting $\mathbf{\Lambda} = (\lambda_{ij})$ denote the rotated factor loading matrix, the varimax rotation maximizes

$$\sum_{j=1}^{m}\{\frac{1}{p}\sum_{i=1}^{p}\lambda_{ij}^4 - \frac{w}{p^2}(\sum_{i=1}^{p}\lambda_{ij}^2)^2\}, \qquad (7)$$

where $w = 1$. The varimax rotation is within the family of orthogonal rotations called the orthomax rotations, which includes quartimax

**Table 4** Varimax-Rotated Factor Pattern Matrix (and standard error in parentheses)

| | *Factor 1* | *Factor 2* | *Factor 3* | *Communality* |
|---|---|---|---|---|
| x1 | 0.143 (0.062) | **0.831** (0.095) | 0.143 (0.075) | 0.732 |
| x2 | 0.135 (0.077) | **0.359** (0.088) | 0.066 (0.074) | 0.151 |
| x3 | 0.256 (0.078) | **0.505** (0.083) | 0.070 (0.076) | 0.325 |
| x4 | **0.833** (0.037) | 0.244 (0.055) | 0.073 (0.055) | 0.759 |
| x5 | **0.794** (0.039) | 0.205 (0.056) | 0.173 (0.057) | 0.702 |
| x6 | **0.780** (0.040) | 0.243 (0.059) | 0.069 (0.058) | 0.672 |
| x7 | 0.166 (0.060) | −0.013 (0.074) | **0.743** (0.077) | 0.580 |
| x8 | −0.013 (0.058) | 0.240 (0.082) | **0.794** (0.074) | 0.688 |
| x9 | 0.188 (0.066) | 0.469 (0.088) | **0.511** (0.078) | 0.517 |

**Table 5** Promax Rotation: Factor Pattern Matrix (and standard error in parentheses)

| | *Factor 1* | *Factor 2* | *Factor 3* | *Communality* |
|---|---|---|---|---|
| x1 | −0.027 (0.041) | **0.875** (0.102) | −0.022 (0.064) | 0.732 |
| x2 | 0.068 (0.094) | **0.359** (0.109) | −0.011 (0.088) | 0.151 |
| x3 | 0.169 (0.093) | **0.494** (0.107) | −0.044 (0.078) | 0.325 |
| x4 | **0.852** (0.044) | 0.061 (0.054) | −0.032 (0.044) | 0.759 |
| x5 | **0.809** (0.046) | 0.007 (0.056) | 0.086 (0.050) | 0.702 |
| x6 | **0.794** (0.047) | 0.074 (0.060) | −0.032 (0.049) | 0.672 |
| x7 | 0.121 (0.054) | −0.193 (0.060) | **0.785** (0.083) | 0.580 |
| x8 | −0.131 (0.043) | 0.126 (0.086) | **0.803** (0.087) | 0.688 |
| x9 | 0.065 (0.070) | 0.388 (0.107) | **0.440** (0.092) | 0.517 |

**Table 6** Promax Rotation: Factor Structure Matrix

| | *Factor 1* | *Factor 2* | *Factor 3* |
|---|---|---|---|
| x1 | 0.341 (0.074) | **0.855** (0.084) | 0.308 (0.095) |
| x2 | 0.219 (0.086) | **0.384** (0.083) | 0.146 (0.086) |
| x3 | 0.369 (0.083) | **0.549** (0.072) | 0.189 (0.076) |
| x4 | **0.870** (0.033) | 0.412 (0.078) | 0.208 (0.086) |
| x5 | **0.834** (0.037) | 0.385 (0.078) | 0.295 (0.085) |
| x6 | **0.817** (0.037) | 0.400 (0.075) | 0.198 (0.082) |
| x7 | 0.239 (0.082) | 0.162 (0.077) | **0.741** (0.074) |
| x8 | 0.127 (0.082) | 0.380 (0.088) | **0.818** (0.068) |
| x9 | 0.343 (0.072) | 0.586 (0.079) | **0.607** (0.076) |

**Table 7** Interfactor Correlation Matrix With Promax Rotation

| | *Factor 1* | *Factor 2* | *Factor 3* |
|---|---|---|---|
| Factor 1 | 1.000 | | |
| Factor 2 | 0.427 (0.076) | 1.000 | |
| Factor 3 | 0.254 (0.089) | 0.386 (0.084) | 1.000 |

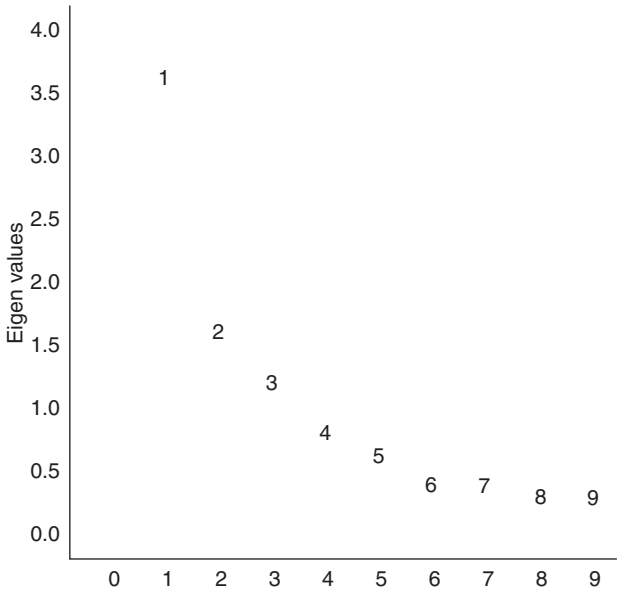



**Figure 1** Scree Plot of Eigenvalues of Sample Correlation Matrix

rotation (with $w = 0$) and equamax criterion (with $w = m/2$).

Among the oblique rotations, the promax rotation is most well-known. The promax rotation is a variant of Procrustes rotation in which a simple structure is achieved by minimizing the distance between the factor pattern matrix and the target matrix using the least-squares method. The promax rotation creates the target matrix using the varimax-rotated factor loading matrix. More specifically, the varimax solution is first normalized, and then each element is accentuated by raising to the power of either three or four while retaining the sign. Other oblique rotations include oblimin rotations that minimize

$$\sum_{j\neq k}^{m}\left\{\sum_{i=1}^{p}\lambda_{ij}^{2}\lambda_{ik}^{2} - \frac{w}{p}\left(\sum_{i=1}^{p}\lambda_{ij}^{2}\right)\left(\sum_{i=1}^{p}\lambda_{ik}^{2}\right)\right\} \quad (8)$$

with the weight $w$ ranging between 0 and 1.

After rotation, the meaning of each factor is identified using the variables on which it has substantial or significant loadings. The significance of loadings can be judged using the ratio of loading over its standard error (SE). The SE option in SAS factor procedure gives outputs of standard errors for factor loadings with various rotations under the normality assumption. Formulas for SE with non-normally distributed or missing data also exist, but not in standard software.

## Factor Scores

It is of interest to know the predicted score for $m$ factors for each observation using the EFA model. Two factor score predictors/estimators are well-known. One is the Bartlett estimator

$$\hat{f}_B = (\mathbf{\Lambda}'\mathbf{\Psi}^{-1}\mathbf{\Lambda})^{-1}\mathbf{\Lambda}'\mathbf{\Psi}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}), \quad (9)$$

which is conditionally unbiased (that is, the expected value of $\hat{f}$ given $f$ is equal to $f$). The other estimator, commonly called the regression estimator,

$$\hat{f}_R = \mathbf{\Phi}\mathbf{\Lambda}'\mathbf{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}), \quad (10)$$

is conditionally biased, but the trace of its mean squared error matrix is the smallest. There are also other different factor score estimators.

## Illustration

To illustrate EFA, nine variables were adopted from the original 26-variable test as reported by Karl J. Holzinger and Frances Swineford in 1939. The nine variables are: ($x_1$) Visual perception, ($x_2$) Cubes, ($x_3$) Flags, ($x_4$) Paragraph comprehension, ($x_5$) Sentence completion, ($x_6$) Word meaning, ($x_7$) Addition, ($x_8$) Counting dots, and ($x_9$) Straight-curved capitals. The sample size is 145, and the sample correlation matrix is given in Table 1. The sample correlation matrix was analyzed with the ML estimation method using the factor procedure in the statistical software SAS version 9.1.3. The SAS code is listed in Table 2. Table 3 contains the eigenvalues of the sample correlation matrix, and the scree plot is given in Figure 1. The first three eigenvalues (3.557, 1.579, and 1.156) were greater than 1. The LRT statistic with Bartlett correction for a three-factor model is 3.443, corresponding to a $p$ value of .992 when compared against the chi-square distribution with 12 degrees of freedom; the LRT statistic for a two-factor model is 49.532, corresponding to a $p$ value of .0002 when compared against the chi-square distribution with 19 degrees of freedom. Thus, both the Kaiser–Guttman criterion and the LRT statistic suggest that a three-factor solution is adequate, and the subsequent analysis was done assuming the three-factor model. After the initial solution, the factor loadings were rotated with the varimax rotation. The varimax-rotated factor loadings are shown in Table 4, where Factor 1 has high loadings on Variables 4 through 6, Factor 2 has high loadings on Variables 1 through 3, and Factor 3 has high loadings on Variables 7 through 9. Communalities are higher than 0.5 except for Variables 2 (0.151) and 3 (0.325). The factor pattern matrix for the promax rotation (Table 5) is almost identical to the factor loading matrix for the varimax rotation. So is the factor structure matrix (Table 6). Thus, it seems obvious that Factor 1 measures language ability, Factor 2 measures visual ability, and Factor 3 measures speed. The correlation between Factors 1 and 2 is 0.427, the correlation between Factors 1 and 3 is 0.254, and the correlation between Factors 2 and 3 is 0.386 (see Table 7).

*Kentaro Hayashi and Ke-Hai Yuan*

***See also*** Confirmatory Factor Analysis; Item Analysis; Latent Variable; Principal

Components Analysis; Structural Equation Modeling

**Further Readings**

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
Gorsuch, R. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
Harmann, H. H. (1976). *Modern factor analysis* (3rd ed.). Chicago: University of Chicago Press.
Hatcher, L. (1994). *A step-by-step approach to using the SAS system for factor analysis and structural equation modeling*. Cary, NC: SAS Institute.
Holzinger, K. J., & Swineford, F. A. (1939). *A study in factor analysis: The stability of a bifactor solution* (Supplementary Educational Monograph 48). Chicago: University of Chicago Press.
Hoyle, R. H., & Duvall, J. L. (2004). Determining the number of factors in exploratory and confirmatory factor analysis. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 301–315). Thousand Oaks, CA: Sage.
Mulaik, S. A. (1972). *The foundations of factor analysis*. New York: McGraw-Hill.
Pett, M. A., Lackey, N. R., & Sullivan, J. J. (2003). *Making sense of factor analysis*. Thousand Oaks, CA: Sage.
Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology, 15,* 201–293.
Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago: University of Chicago Press.
Yanai, H., & Ichikawa, M. (2007). Factor analysis. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 257–296). Amsterdam: Elsevier.
Yuan, K.-H., Marshall, L. L., & Bentler, P. M. (2002). A unified approach to exploratory factor analysis with missing data, nonnormal data, and in the presence of outliers. *Psychometrika, 67,* 95–122.

## Ex Post Facto Study

Ex post facto study or after-the-fact research is a category of research design in which the investigation starts after the fact has occurred without interference from the researcher. The majority of social research, in contexts in which it is not possible or acceptable to manipulate the characteristics of human participants, is based on ex post facto research designs. It is also often applied as a substitute for true experimental research to test hypotheses about cause-and-effect relationships or in situations in which it is not practical or ethically acceptable to apply the full protocol of a true experimental design. Despite studying facts that have already occurred, ex post facto research shares with experimental research design some of its basic logic of inquiry.

Ex post facto research design does not include any form of manipulation or measurement before the fact occurs, as is the case in true experimental designs. It starts with the observation and examination of facts that took place naturally, in the sense that the researcher did not interfere, followed afterward by the exploration of the causes behind the evidence selected for analysis. The researcher takes the dependent variable (the fact or effect) and examines it retrospectively in order to identify possible causes and relationships between the dependent variable and one or more independent variables. After the deconstruction of the causal process responsible for the facts observed and selected for analysis, the researcher can eventually adopt a prospective approach, monitoring what happens after that.

Contrary to true experimental research, ex post facto research design looks first to the effects (dependent variable) and tries afterward to determine the causes (independent variable). In other words, unlike experimental research designs, the independent variable has already been applied when the study is carried out, and for that reason, it is not manipulated by the researcher. In ex post facto research, the control of the independent variables is made through statistical analysis, rather than by control and experimental groups, as is the case in experimental designs. This lack of direct control of the independent variable and the nonrandom selection of participants are the most important differences between ex post facto research and the true experimental research design.

Ex post facto research design has strengths that make it the most appropriate research plan in numerous circumstances; for instance, when it is not possible to apply a more robust and rigorous research design because the phenomenon occurred naturally; or it is not practical to

manipulate the independent variables; or the control of independent variables is unrealistic; or when such manipulation of human participants is ethically unacceptable (e.g., delinquency, illnesses, road accidents, suicide). Instead of exposing human subjects to certain experiments or treatments, it is more reasonable to explore the possible causes after the fact or event has occurred, as is the case in most issues researched in anthropology, geography, sociology, and in other social sciences. It is also a suitable research design for an exploratory investigation of cause-effect relationships or for the identification of hypotheses that can later be tested through true experimental research designs.

It has a number of weaknesses or shortcomings as well. From the point of view of its internal validity, the two main weak points are the lack of control of the independent variables and the nonrandom selection of participants or subjects. For example, its capacity to assess confounding errors (e.g., errors due to history, social interaction, maturation, instrumentation, selection bias, mortality) is unsatisfactory in numerous cases. As a consequence, the researcher may not be sure that all independent variables that caused the facts observed were included in the analysis, or if the facts observed would not have resulted from other causes in different circumstances, or if that particular situation is or is not a case of reverse causation. It is also open to discussion whether the researcher will be able to find out if the independent variable made a significant difference or not in the facts observed, contrary to the true experimental research design, in which it is possible to establish if the independent variable is the cause of a given fact or event. Therefore, from the point of view of its internal validity, ex post facto research design is less persuasive to determine causality compared to true experimental research designs. Nevertheless, if there is empirical evidence flowing from numerous case studies pointing to the existence of a causal relationship, statistically tested, between the independent and dependent variables selected by the researcher, it can be considered sound evidence in support of the existence of a causal relationship between these variables. It has also a number of weaknesses from the point of view of its external validity, when samples are not randomly selected (e.g., nonprobabilistic samples: convenient samples, snowball samples), which limit the possibility of statistical inference. For that reason, findings in ex post facto research design cannot, in numerous cases, be generalized or looked upon as being statistically representative of the population.

In sum, ex post facto research design is widely used in social as well as behavioral and biomedical sciences. It has strong points that make it the most appropriate research design in a number of circumstances as well as limitations that make it weak from the point of view of its internal and external validity. It is often the best research design that can be used in a specific context, but it should be applied only when a more powerful research design cannot be employed.

*Carlos Nunes Silva*

***See also*** Cause and Effect; Control Group; Experimental Design; External Validity; Internal Validity; Nonexperimental Designs; Pre-Experimental Design; Quasi-Experimental Design; Research Design Principles

**Further Readings**

Bernard, H. R. (1994). *Research methods in anthropology: Qualitative and quantitative approaches* (2nd ed.). Thousand Oaks, CA: Sage.

Black, T. R. (1999). *Doing quantitative research in the social sciences: An integrated approach to research design, measurement and statistics*. Thousand Oaks, CA: Sage.

Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education*. London: Routledge.

Engel, R. J., & Schutt, R. K. (2005). *The practice of research in social work*. Thousand Oaks, CA: Sage.

Ethridge, M. E. (2002). *The political research experience. Readings and analysis* (3rd ed.). Armonk, NY: M. E. Sharpe.

## External Validity

When an investigator wants to generalize results from a research study to a wide group of people (or a population), he or she is concerned with external validity. A set of results or conclusions

from a research study that possesses external validity can be generalized to a broader group of individuals than those originally included in the study. External validity is relevant to the topic of research methods because scientific and scholarly investigations are normally conducted with an interest in generalizing findings to a larger population of individuals so that the findings can be of benefit to many and not just a few. In the next three sections, the kinds of generalizations associated with external validity are introduced, the threats to external validity are outlined, and the methods to increase the external validity of a research investigation are discussed.

## Two Kinds of Generalizations

Two kinds of generalizations are often of interest to researchers of scientific and scholarly investigations: (a) generalizing research findings to a specific or target population, setting, and time frame; and (b) generalizing findings across populations, settings, and time frames. An example is provided to illustrate the difference between the two kinds. Imagine a new herbal supplement is introduced that is aimed at reducing anxiety in 25-year-old women in the United States. Suppose that a random sample of all 25-year-old women has been drawn that provides a nationally representative sample within known limits of sampling error. Imagine now that the women are randomly assigned to two conditions—one where the women consume the herbal supplement as prescribed, and the other a control group where the women unknowingly consume a sugar pill. The two conditions or groups are equivalent in terms of their representativeness of 25-year-old women. Suppose that after data analysis, the group that consumed the herbal supplement demonstrated lower anxiety than the control group as measured by a paper-and-pencil questionnaire. The investigator can generalize this finding to the average 25-year-old woman in the United States, that is, the target population of the study. Note that this finding can be generalized to the average 25-year-old woman despite possible variations in how differently women in the experimental group reacted to the supplement. For example, a closer analysis of the data might reveal that women in the experimental group who exercised regularly reduced their anxiety more in relation to women who did not; in fact, a closer analysis might reveal that only those women who exercised regularly in addition to taking the supplement reduced their anxiety. In other words, closer data analysis could reveal that the findings do not generalize across all subpopulations of 25-year-old women (e.g., those who do not exercise) even though they do generalize to the overall target population of 25-year-old women.

The distinction between these two kinds of generalizations is useful because generalizing to specific populations is surprisingly more difficult than generalizing across populations because the former typically requires large-scale studies where participants have been selected using formal random sampling procedures. This is rarely achieved in field research, where large-scale studies pose challenges for administering treatment interventions and for high-quality measurement, and participant attrition is liable to occur systematically. Instead, the more common practice is to generalize findings from smaller studies, each with its own sample of convenience or accidental sampling (i.e., a sample that is accrued expediently for the purpose of the research but provides no guarantee that it formally represents a specific target population), across the populations, settings, and time frames associated with the smaller studies. It needs to be noted that individuals in samples of convenience may belong to the target population to which one wishes to generalize findings; however, without formal random sampling, the representativeness of the sample is questionable. According to Thomas Cook and Donald Campbell, an argument can be made for strengthening external validity by means of a greater number of smaller studies with samples of convenience than by a single large study with an initially representative sample. Given the frequency of generalizations across populations, settings, and time frames in relation to target populations, the next section reviews the threats to external validity claims associated with this type of generalization.

## Threats to External Validity

To be able to generalize research findings across populations, settings, and time frames, the investigator needs to have evidence that the research findings are not unique to a single population,

but rather apply to more than one population. One source for this type of evidence comes from examining statistical interactions between variables of interest. For example, in the course of data analysis, an investigator might find that consuming an herbal supplement (experimental treatment) statistically interacts with the activity level of the women participating in the study, such that women who exercise regularly benefit more from the anxiety-reducing effects of the supplement relative to women who do not exercise regularly. What this interaction indicates is that the positive effects of the herbal supplement cannot be generalized equally to all subpopulations of 25-year-old women. The presence of a statistical interaction means that the effect of the variable of interest (i.e., consuming the herbal supplement) changes across levels of another variable (i.e., activity levels of 25-year-old women). In order to generalize the effects of the herbal supplement across subpopulations of 25-year-old women, a statistical interaction cannot be observed between the two variables of interest. Many interactions can threaten the external validity of a study. These are outlined as follows.

### *Participant Selection and Treatment Interaction*

To generalize research findings across populations of interest, it is necessary to recruit participants in an unbiased manner. For example, when recruiting female participants to take part in an herbal supplement study, if the investigator advertises the study predominantly in health food stores and obtains the bulk of participants from this location, then the research findings may not generalize to women who do not visit health food stores. In other words, there may be something unique to those women who visit health food stores and decide to volunteer in the study that may make them more disposed to the effects of a health supplement. To counteract this potential bias, the investigator could systematically advertise the study in other kinds of food stores to test whether the selection of participants from different locations interacts with the treatment. If the statistical interaction is absent, then the investigator can be confident that the research findings are not exclusive to those women who visit health food stores and, possibly, are more susceptible to the effects of an herbal supplement than other women. Thus, recruiting participants from a variety of locations and making participation as convenient as possible should be undertaken.

### *Setting and Treatment Interaction*

Just as the selection of participants can interact with the treatment, so can the setting in which the study takes place. This type of interaction is more applicable to research studies where participants experience an intervention that could plausibly change in effect depending on the context, such as in educational research or organizational psychological investigations. However, to continue with the health supplement example, suppose the investigator requires the participants to consume the health supplement in a laboratory and not in their homes. Imagine that the health supplement produces better results when the participant ingests it at home and produces worse results when the participant ingests it in a laboratory setting. If the investigator varies the settings in the study, it is possible to test the statistical interaction between the setting in which the supplement is ingested and the herbal supplement treatment. Again, the absence of a statistical interaction between the setting and the treatment variable would indicate that the research findings can be generalized across the two settings; the presence of an interaction would indicate that the findings cannot be generalized across the settings.

### *History and Treatment Interaction*

In some cases, the historical time in which the treatment occurs is unique and could contribute to either the presence or absence of a treatment effect. This is a potential problem because it means that whatever effect was observed cannot be generalized to other time frames. For example, suppose that the herbal supplement is taken by women during a week in which the media covers several high-profile optimistic stories about women. It is reasonable for an investigator to inquire whether the positive results of taking an herbal supplement would have been obtained during a less eventful week. One way to test for the interaction between

historical occurrences and treatment is to administer the study at different time frames and to replicate the results of the study.

## Methods to Increase External Validity

If one wishes to generalize research findings to target populations, it is appropriate to outline a sampling frame and select instances so that the sample is representative of the population to which one wishes to generalize within known limits of sampling error. Procedures for how to do this can be found in textbooks on sampling theory. Often, the most representative samples will be those that have been selected randomly from the population of interest. This method of *random sampling for representativeness* requires considerable resources and is often associated with large-scale studies. After participants have been randomly selected from the population, participants can then be randomly assigned to experimental groups.

Another method for increasing external validity involves *sampling for heterogeneity.* This method requires explicitly defining target categories of persons, settings, and time frames to ensure that a broad range of instances from within each category is represented in the design of the study. For example, an educational researcher interested in testing the effects of a mathematics intervention might design the study to include boys and girls from both public and private schools located in small rural towns and large metropolitan cities. The objective would then be to test whether the intervention has the same effect in all categories (e.g., whether the mathematics intervention leads to the same effect in boys and girls, public and private schools, and rural and metropolitan areas). Testing for the effect in each of the categories requires a sufficiently large sample size in each of the categories. Deliberate sampling for heterogeneity does not require random sampling at any stage in the design, so it is usually viable to implement in cases where investigators are limited by resources and in their access to participants. However, deliberate sampling does not allow one to generalize from the sample to any formally specified population. What deliberate sampling does allow one to conclude is that an effect has or has not been obtained within a specific range of categories of persons, settings, and times. In other words, one can claim that "in at least one sample of boys and girls, the mathematics intervention had the effect of increasing test scores."

There are other methods to increase external validity, such as the *impressionistic modal instance model,* where the investigator samples purposively for specific types of instances. Using this method, the investigator specifies the category of person, setting, or time to which he or she wants to generalize and then selects an instance of each category that is impressionistically similar to the category mode. This method of selecting instances is most often used in consulting or project evaluation work where broad generalizations are not required. The most powerful method for generalizing research findings, especially if the generalization is to a target population, is random sampling for representativeness. The next most powerful method is random sampling for heterogeneity, with the method of impressionistic modal instance being the least powerful. The power of the model decreases as the natural assortment of individuals in the sample dwindles. However, practical concerns may prevent an investigator from using the most powerful method.

*Jacqueline P. Leighton*

***See also*** Inference: Deductive and Inductive; Interaction; Research Design Principles; Sampling; Theory

### Further Readings

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Boston: Houghton Mifflin.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin.

Tebes, J. K. (2000). External validity and scientific psychology. *American Psychologist, 55*(12), 1508–1509.

# F

## Face Validity

Face validity is a test of internal validity. As the name implies, it asks a very simple question: "On the face of things, do the investigators reach the correct conclusions?" It requires investigators to step outside of their current research context and assess their observations from a commonsense perspective. A typical application of face validity occurs when researchers obtain assessments from current or future individuals who will be directly affected by programs premised on their research findings. An example of testing for face validity is the assessment of a proposed new patient tracking system by obtaining observations from local community health care providers who will be responsible for implementing the program and getting feedback on how they think the new program may work in their centers.

What follows is a brief discussion on how face validity fits within the overall context of validity tests. Afterward, documentation of face validity's history is reviewed. Here, early criticisms of face validity are addressed that set the stage for how and why the test returned as a valued assessment. This discussion of face validity concludes with some recent applications of the test.

### The Validity of Face Validity

To better understand the value and application of face validity, it is necessary to first set the stage for what validity is. Validity is commonly defined as a question: "To what extent do the research conclusions provide the correct answer?" In testing the validity of research conclusions, one looks at the relationship of the purpose and context of the research project to the research conclusions. Validity is determined by testing (questions of validity) research observations against what is already known in the world, giving the phenomenon that researchers are analyzing the chance to prove them wrong. All tests of validity are context-specific and are not an absolute assessment. Tests of validity are divided into two broad realms: external validity and internal validity. Questions of external validity look at the generalizability of research conclusions. In this case, observations generated in a research project are assessed on their relevance to other, similar situations. Face validity falls within the realm of internal validity assessments. A test of internal validity asks if the researcher draws the correct conclusion based on the available data. These types of assessments look into the nuts-and-bolts of an investigation (for example, looking for sampling error or researcher bias) to see if the research project was legitimate.

### History of Face Validity

For all of its simplicity, the test for face validity has had an amazing and dramatic past that, until recently, has re-emerged as a valued and respected test of validity. In its early applications, face validity was used by researchers as

**Table 1** Debate Over Face Validity

| *Point of Contention* | *Two Sides of the Debate on Face Validity* | |
|---|---|---|
| Is face validity a valid test? | **Face validity is not a valid test.** Several other, more technically advanced tests of validity provide a more detailed assessment of research than what is learned from a face validity assessment. | **Face validity is a valid test.** Face validity asks a simple question: Do the research findings make sense? No other validity test that assesses research findings is based on common sense. Face validity fills a gap in internal validity tests. |
| Is face validity a stand-alone test, or is it another shade of content validity? | **Content validity and face validity are related.** Face validity is a watered-down version of content validity because it asks roughly the same question: Do the identified research variables closely fit what is known about the research topic? When you have content validity, you automatically have face validity. | **Face validity is a stand-alone test.** Face validity focuses on the commonsense *appearance* of validity in the research findings. Content validity focuses on the *fit* of the defined content of variables to what is known about the research topic. When you have content validity (fit), you may not have face validity (appearance). |
| Is face validity only for experts? | **Only experts can conduct face validity tests.** Face validity is for experts only because laypeople do not have a true understanding of research methods and tests of validity. | **Laypersons are important for tests of face validity.** Laypersons provide valuable insights on the applicability of the research findings based on their experiential perspectives. |

a first-step assessment, in concert with other tests, to assess the validity of an analysis. During the 1940s and 1950s, face validity was used by psychologists when they were in the early stages of developing tests for use in selecting industrial and military personnel. It was soon widely used by many different types of researchers in different types of investigations, resulting in confusion on what actually constituted face validity. Quickly, the confusion over the relevance of face validity gave way to its being rejected by researchers in the 1960s, who took to new and more complex tests of validity.

## Early Debate Surrounding Face Validity

Discussions surrounding face validity were revived in 1985 by Baruch Nevo's seminal article "Face Validity Revisited," which focused on clearing up some of the confusion surrounding the test and challenging researchers to take another, more serious look at face validity's applications. Building on Nevo's research, three questions can be distinguished in the research validity literature that have temporarily prevented face validity from getting established as a legitimate test of validity (see Table 1).

The first question regarding face validity is over the legitimacy of the test itself. Detractors argue that face validity is insignificant because its observations are not based on any verifiable testing procedure yielding only rudimentary observations about a study. Face validity does not require a systematic method in the obtaining of face validity observations. They conclude that the only use for face validity observations is for public relations statements.

Advocates for face validity see that face validity provides researchers with the opportunity for commonsense testing of research results: "After the investigation is completed and all the tests of validity and reliability are done, does this study make sense?" Here, tests of face validity allow investigators a new way to look at their conclusions to make sure they see the forest for the trees, with the forest being common sense and the trees being all of the different tests of validity used in documenting the veracity of their study.

The second question confuses the value of face validity by blurring the applications of face validity with content validity. The logic here is that both tests of validity are concerned with content and the representativeness of the study. Content validity is the extent to which the items identified in the study reflect the domain of the concept being measured. Because content validity and face validity both look at the degree to which the intended range of meanings in the concepts of the study appear to be covered, once a study has content validity, it will automatically have face validity. After testing for content validity, there is no real need to test for face validity.

The other side to this observation is that content validity should not be confused with face validity because they are completely different tests. The two tests of validity are looking at different parts of the research project. Content validity is concerned with the relevance of the identified research variables within a proposed research project, whereas face validity is concerned with the relevance of the overall completed study. Face validity looks at the overall commonsense assessment of a study. In addition to the differences between the two tests of validity in terms of what they assess, other researchers have identified a sequential distinction between content validity and face validity. Content validity is a test that should be conducted before the data-gathering stage of the research project is started, whereas face validity should be applied after the investigation is carried out. The sequential application of the two tests is intuitively logical because content validity focuses on the appropriateness of the identified research items before the investigation has started, whereas face validity is concerned with the overall relevance of the research findings after the study has been completed.

The third question surrounding face validity asks a procedural question: Who is qualified to provide face validity observations—experts or laypersons? Proponents for the "experts-only" approach to face validity believe that experts who have a substantive knowledge about a research topic and a good technical understanding of tests of validity provide constructive insights from outside of the research project. In this application of face validity, experts provide observations that can help in the development and/or fine-tuning of research projects. Laypersons lack technical research skills and can provide only impressionistic face validity observations, which are of little use to investigators.

Most researchers now see that the use of experts in face validity assessments is more accurately understood as being a test of content validity because they provide their observations at the start or middle of a research project, and face validity focuses on assessing the relevance of research conclusions. Again, content validity should be understood sequentially in relation to face validity, with the former being used to garner expert observations on the relevance of research variables in the earlier parts of the investigation from other experts in the field, and face validity should come from laypersons for their commonsense assessment at the completion of the research project.

The large-scale vista that defines face validity, defines the contribution this assessment provides to the research community, also provides its Achilles heel. Face validity lacks the depth, precision, and rigor of inquiry that comes with both internal and external validity tests. For example, in assessing the external validity of a survey research project, one can precisely look at the study's sample size to determine if it has a representative sample of the population. The only question face validity has for a survey research project is a simple one: "Does the study make sense?" For this reason, face validity can never be a stand-alone test of validity.

## The Re-Emergence of Face Validity

The renewed interest in face validity is part of the growing research practice of integrating laypersons' nontechnical, one-of-a-kind insights into the evaluation of applied research projects. Commonly known as obtaining an emic viewpoint, testing for face validity provides the investigator the opportunity to learn what many different people affected by a proposed program already know about a particular topic. The goal in this application of face validity is to include the experiential perspectives of people affected by research projects in their assessment of what causes events to happen, what the effects of the study in the community may be, and what specific words or events mean in the community.

The following examples show how researchers use face validity assessments in very different contexts, but share the same goal: obtaining a commonsense assessment from persons affected by research conclusions. Michael Quinn Patton is widely recognized for his use of "internal evaluators" to generate face validity observations in the evaluation of programs. In the Hazelden Foundation of Minnesota case study, he describes his work in providing annual evaluations based on the foundation's data of tracking clients who go through its program. At the completion of the annual evaluation, a team of foundation insider evaluators then participates in the evaluation by assessing the data and conclusions made in the reports.

Face validity assessments are commonly used in applied research projects that include the fields of community development, planning, public policy, and macro social work. In planning, face validity observations are obtained during scheduled public hearings throughout the planning process. The majority of planning research is based on artificial constructs of reality that allow planners to understand complex, multivariable problems (e.g., rush-hour traffic). One of the reasons that planners incorporate citizen input into the planning process is that it allows them to discover the "inside perspective" from the community on how their research and proposed plans may affect their day-to-day lives. A street-widening project in Lincoln, Nebraska, is one example of how a city used face validity in its planning process. A central traffic corridor was starting to experience higher levels of rush-hour congestion as the result of recent growth on the city's edge. Knowing that simply widening the street to accommodate more vehicles could affect area businesses adversely, city planners met with local store owners to get their face validity observations of how the street affected their daily operations. Armed with traffic data and face validity observations of local store owners, the city was able to plan a wider street that took into account both traffic commuters' and area businesses' experiences with the street.

*John Gaber*

***See also*** Applied Research; Planning Research; "Validity"

**Further Readings**

Nevo, B. (1985). Face validity revisited. *Journal of Educational Measurement, 22*(4), 287–293.
Patton, M. Q. (1997). *Utilization-focused evaluation* (3rd ed.). Thousand Oaks, CA: Sage.
Pike, K. (1967). *Language in relation to a unified theory of the structure of human behavior.* Paris: Mouton & Co.

# Factorial Design

A factorial design contains two or more independent variables and one dependent variable. The independent variables, often called *factors,* must be categorical. Groups for these variables are often called *levels.* The dependent variable must be continuous, measured on either an interval or a ratio scale.

Suppose a researcher is interested in determining if two categorical variables (treatment condition and gender) affect a continuous variable (achievement). The researcher decides to use a factorial design because he or she wants to examine population group means. A factorial analysis of variance will allow him or her to answer three questions. One question concerns the main effect of treatment: Do average achievement scores differ significantly across treatment conditions? Another question concerns the main effect of gender: Does the average achievement score for females differ significantly from the average achievement score for males? The final question refers to the interaction effect of treatment condition and gender: Is the effect of treatment condition on achievement the same for both genders?

This entry first describes how to identify factorial designs and their advantages. Next, analysis and interpretation of factorial designs, including follow-up analyses for significant results, are discussed. A short discussion on the importance of effect size concludes the entry.

## Identification

One way to identify factorial designs is by the number of factors involved. Although there is no limit to the number of factors, two-factor and three-factor designs are most common. Occasionally,

a researcher will use a four-factor design, but these situations are extremely rare. When a study incorporates a large number of factors, other designs are considered, such as regression.

Another way to identify factorial designs is by the number of levels for each factor. The simplest design is a 2 × 2, which represents two factors, both of them having two levels. A 3 × 4 design also has two factors, but one factor has three levels (e.g., type of reward: none, food, money) and the other factor has 4 levels (e.g., age: 6–8 years, 9–11 years, 12–14 years, 15–16 years). A 2 × 2 × 3 design has three factors; for example, gender (2 levels: male, female), instructional method (2 levels: traditional, computer-based), and ability (3 levels: low, average, high).

In a factorial design, each level of a factor is paired with each level of another factor. As such, the design includes all combinations of the factors' levels, and a unique subset of participants is in each combination. Using the 3 × 4 example in the previous paragraph, there are 12 cells or subsets of participants. If a total of 360 participants were included in the study and group sample sizes were equal, then 30 young children (ages 6 to 8) would receive no reward for completing a task, a different set of 30 young children (ages 6 to 8) would receive food for completing a task, and yet a different set of 30 young children (ages 6 to 8) would receive money for completing a task. Similarly, unique sets of 30 children would be found in the 9–11, 12–14, and 15–16 age ranges.

This characteristic separates factorial designs from other designs that also involve categorical independent variables and continuous dependent variables. For instance, a repeated measures design requires the same participant to be included in more than one level of an independent variable. If the 3 × 4 example was changed to a repeated measures design, then each participant would be exposed to tasks involving the three different types of rewards: none, food, and money.

## Advantages

Factorial designs have several advantages. First, they allow for a broader interpretation of results. If a single-factor design was used to examine treatments, the researcher could generalize results only to the characteristics of the particular group of participants chosen, whereas if one or two additional factors, such as gender or age, are included in the design, then the researcher can examine differences between these specific subsets of participants. Another advantage is that the simultaneous effect of the factors operating together can be tested. By examining the interaction between treatment and age, the researcher can determine whether the effect of treatment is dependent on age. The youngest participant group may show higher scores when receiving Treatment A, whereas the oldest participant group may show higher scores when receiving Treatment B.

A third advantage of factorial designs is that they are more parsimonious, efficient, and powerful than an examination of each factor in a separate analysis. The principle of parsimony refers to conducting one analysis to answer all questions rather than multiple analyses. Efficiency is a related principle. Using the most efficient design is desirable, meaning the one that produces the most precise estimate of the parameters with the least amount of sampling error. When additional factors are added to a design, the error term can be greatly reduced. A reduction of error also leads to more powerful statistical tests. A factorial design requires fewer participants in order to achieve the same degree of power as in a single-factor design.

## Analysis and Interpretation

The statistical technique used for answering questions from a factorial design is the analysis of variance (ANOVA). A factorial ANOVA is an extension of a one-factor ANOVA. A one-factor ANOVA involves one independent variable and one dependent variable. The *F*-test statistic is used to test the null hypothesis of equality of group means. If the dependent variable is reaction time and the independent variable has three groups, then the null hypothesis states that the mean reaction times for Groups 1, 2, and 3 are equal. If the *F* test leads to rejection of the null hypothesis, then the alternative hypothesis is that at least one pair of the group mean reaction times is not equal. Follow-up analyses are necessary to determine which pair or pairs of means are unequal.

Additional null hypotheses are tested in a factorial ANOVA. For a two-factor ANOVA, there are three null hypotheses. Two of them assess main

**Table 1** Sample Means for the Mathematics Instruction Study

| | *Method A Problem Solving* | *Method B Traditional* | *Marginal Means for Gender* |
|---|---|---|---|
| Males | $\overline{X}_{A,\ males} = 55.0$ | $\overline{X}_{B,\ males} = 40.0$ | $\overline{X}_{males} = 47.5$ |
| Females | $\overline{X}_{A,\ females} = 50.0$ | $\overline{X}_{B,\ females} = 35.0$ | $\overline{X}_{females} = 42.5$ |
| Marginal Means for Method | $\overline{X}_{A} = 52.5$ | $\overline{X}_{B} = 37.5$ | $\overline{X}_{overall} = 45.0$ |

effects, that is, the independent effect of each independent variable on the dependent variable. A third hypothesis assesses the interaction effect of the two independent variables on the dependent variable. For a three-factor ANOVA, there are seven null hypotheses: (a) three main-effect hypotheses, one for each independent variable; (b) three two-factor interaction hypotheses, one for each unique pair of independent variables; and (c) one three-factor interaction hypothesis that examines whether a two-factor interaction is generalizable across levels of the third factor. It is important to note that each factorial ANOVA examines only one dependent variable. Therefore, it is called a univariate ANOVA. When more than one dependent variable is included in a single procedure, a multivariate ANOVA is used.

### *Model Assumptions*

Similar to one-factor designs, there are three model assumptions for a factorial analysis: normality, homogeneity of variance, and independence. First, values on the dependent variable within each population group must be normally distributed around the mean. Second, the population variances associated with each group in the study are assumed to be equal. Third, one participant's value on the dependent variable should not be influenced by any other participant in the study. Although not an assumption per se, another requirement of factorial designs is that each subsample should be a random subset from the population. Prior to conducting statistical analysis, researchers should evaluate each assumption. If assumptions are violated, the researcher can either (a) give evidence that the inferential tests are robust and the probability statements remain valid or (b) account for the violation by transforming variables, use statistics that adjust for the violation, or use nonparametric alternatives.

### *Example of Math Instruction Study*

The following several paragraphs illustrate the application of factorial ANOVA for an experiment in which a researcher wants to compare effects of two methods of math instruction on math comprehension. Method A involves a problem-solving and reasoning approach, and Method B involves a more traditional approach that focuses on computation and procedures. The researcher also wants to determine whether the methods lead to different levels of math comprehension for male versus female students. This is an example of a 2 × 2 factorial design. There are two independent variables: method and gender. Each independent variable has two groups or levels. The dependent variable is represented by scores on a mathematics comprehension assessment. Total sample size for the study is 120, and there are 30 students in each combination of gender and method.

#### Matrix of Sample Means

Before conducting the ANOVA, the researcher examined a matrix of sample means. There are three types of means in a factorial design: cell means, marginal means, and an overall (or grand) mean. In a 2 × 2 design, there are four cell means, one for each unique subset of participants. Table 1 shows that the 30 male students who received Method A had an average math comprehension score of 55. Males in Method B had an average score of 40. Females' scores were lower but had the same pattern across methods as males' scores. The 30 female students in Method A had an average score of 50, whereas the females in Method B had a score of 35.

The second set of means is called marginal means. These means represent the means for all students in one group of one independent variable. Gender marginal means represent the mean of all 60 males (47.5) regardless of which method they

**Table 2** ANOVA Summary Table for the Mathematics Instruction Study

| | *SS* | *df* | *MS* | *F* | *p* |
|---|---|---|---|---|---|
| Method | 6157.9 | 1 | 6157.9 | 235.8 | < .001 |
| Gender | 879.0 | 1 | 879.0 | 33.7 | < .001 |
| Method × Gender | 1.9 | 1 | 1.9 | 0.1 | .790 |
| Within (error) | 3029.3 | 116 | 26.1 | | |
| Total | 10068.2 | 119 | | | |

received, and likewise the mean of all 60 females (42.5). Method marginal means represent the mean of all 60 students who received Method A (52.5) regardless of gender, and the mean of 60 students who received Method B (37.5). Finally, the overall mean is the average score for all 120 students (45.0) regardless of gender or method.

### The *F* Statistic

An *F*-test statistic determines whether each of the three null hypotheses in the two-factor ANOVA should be rejected or not rejected. The concept of the *F* statistic is similar to that of the *t* statistic for testing the significance of two group means. It is a ratio of two values. The numerator of the *F* ratio is the variance that can be attributed to the observed differences between the group means. The denominator is the amount of variance that is "left over," that is, the amount of variance due to differences among participants within groups (or error). Therefore, the *F* statistic is a ratio between two variances—variance attributable "between" groups and variance attributable "within" groups. Is the between-groups variance larger than the within-groups variance? The larger it is, the larger the *F* statistic. The larger the *F* statistic, the more likely it is that the null hypothesis will be rejected. The observed *F* (calculated from the data) is compared to a critical *F* at a certain set of degrees of freedom and significance level. If the observed *F* is larger than the critical *F*, then the null hypothesis is rejected.

### Partitioning of Variance

Table 2 shows results from the two-factor ANOVA conducted on the math instruction study. An ANOVA summary table is produced by statistical software programs and is often presented in research reports. Each row identifies a portion of variation. Within rows, there are several elements: sum of squares (SS), degrees of freedom (*df*), mean square (MS), *F* statistic, and significance level (*p*).

The last row in the table represents the total variation in the data set. SS(total) is obtained by determining the deviation between each individual raw score and the overall mean, squaring the deviations, and obtaining the sum. The other rows partition this total variation into four components. Three rows represent between variation, and one represents error variation. The first row in Table 2 shows the between source of variation due to method. To obtain SS(method), each method mean is subtracted from the overall mean, the deviations are squared, multiplied by the group sample size, and then summed. The degrees of freedom for method is the number of groups minus 1. For the between source of variation due to gender, the sum of squares is found in a similar way by subtracting the gender group means from the overall mean. The third row is the between source of variation accounted for by the interaction between method and gender. The sum of squares is the overall mean minus the effects of method and gender plus the individual cell effect. The degrees of freedom are the product of the method and gender degrees of freedom. The fourth row represents the remaining unexplained variation not accounted for by the two main effects and the interaction effect. The sum of squares for this error variation is obtained by finding the deviation between each individual raw score and the mean of the subgroup to which it belongs, squaring that deviation, and then summing all deviations. Degrees of freedom for the between and within sources of variation add up to the *df*(total), which is the total number of individuals minus 1.

As mentioned earlier, mean square represents variance. The mean square is calculated in the same way as the variance for any set of data. Therefore, the mean square in each row of Table 2 is the ratio between SS and *df*. Next, in order to make the decision about rejecting or not rejecting each null hypothesis, the *F* ratio is calculated. Because the *F* statistic is the ratio of between to within variance, it is simply obtained by dividing the mean square for each between source by the mean square for error. Finally, the *p* values in Table 2 represent the significance level for each null hypothesis tested.

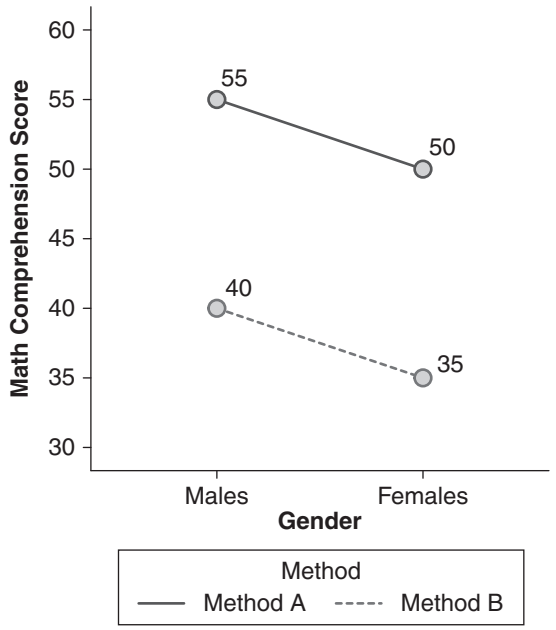



**Figure 1** Plot of Nonsignificant Interaction of Method by Gender

### Interpreting the Results

Answering the researcher's questions requires examining the $F$ statistic to determine if it is large enough to reject the null hypothesis. A critical value for each test can be found in a table of critical $F$ values by using the $df$(between) and $df$(within) and the alpha level set by the researcher. If the observed $F$ is larger than the critical $F$, then the null hypothesis is rejected. Alternatively, software programs provide the observed $p$ value for each test. The null hypothesis is rejected when the $p$ value is less than the alpha level.

In this study, the null hypothesis for the main effect of method states that the mean math comprehension score for all students in Method A equals the mean score for all students in Method B, regardless of gender. Results indicate that the null hypothesis is rejected, $F(1, 116) = 235.8$, $p < .001$. There is a significant difference in math comprehension for students who received the two different types of math instruction. Students who experienced the problem-solving and reasoning approach had a higher mean score (52.5) than students with the traditional approach (37.5). The null hypothesis for the main effect of gender states that the mean score for all males equals the mean score for all females, regardless of the method they received. Results shows that the null hypothesis is rejected, $F(1, 116) = 33.7$, $p < .001$. The mean score for all males (47.5) is higher than the mean score for all females (42.5), regardless of method. Finally, results show no significant interaction between method and gender, $F(1, 116) = .1$, $p = .790$, meaning that the difference in mean math scores for Method A versus Method B is the same for males and females. For both genders, the problem-solving approach produced a higher mean score than the traditional approach. Figure 1 shows the four cell means plotted on a graph. The lines are parallel, indicating no interaction between method and gender.

### A Further Look at Two-Factor Interactions

When the effect of one independent variable is constant across the levels of the other independent variable, there is no significant interaction. Pictorially, the lines on the graph are parallel. When the lines are significantly nonparallel, then there is an interaction between the two independent variables. Generally, there are two types of significant interactions: ordinal and disordinal. Figure 2 illustrates an ordinal interaction. Suppose that the math instruction study showed a significant interaction effect, indicating that the effect of method is not the same for the two genders. The cell means on the graph show nonparallel lines that do not intersect. Although the means for Method A are higher than those for Method B for both genders, there is a 15-point difference in the method means for males, but only a 5-point difference in the method means for females.

Figure 3 illustrates a disordinal interaction in which the lines intersect. In this scenario, Method A produced a higher mean score for males, and Method B produced a higher mean score for females. The magnitude of the difference in gender means for each method is the same, but the direction is different. Method A mean minus Method B mean for males was $+15$, whereas the difference between Method A and Method B means for females was $-15$.

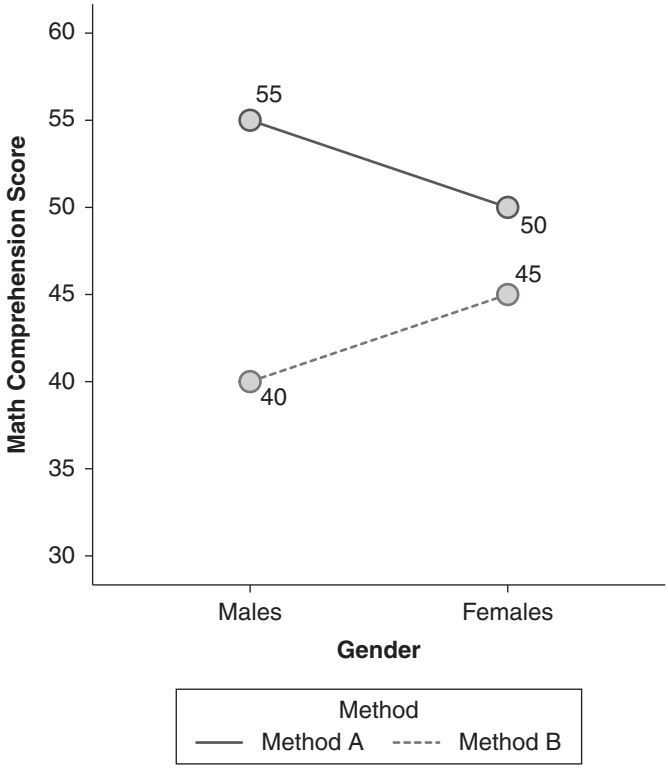



**Figure 2** Plot of Significant, Ordinal Interaction of Method by Gender

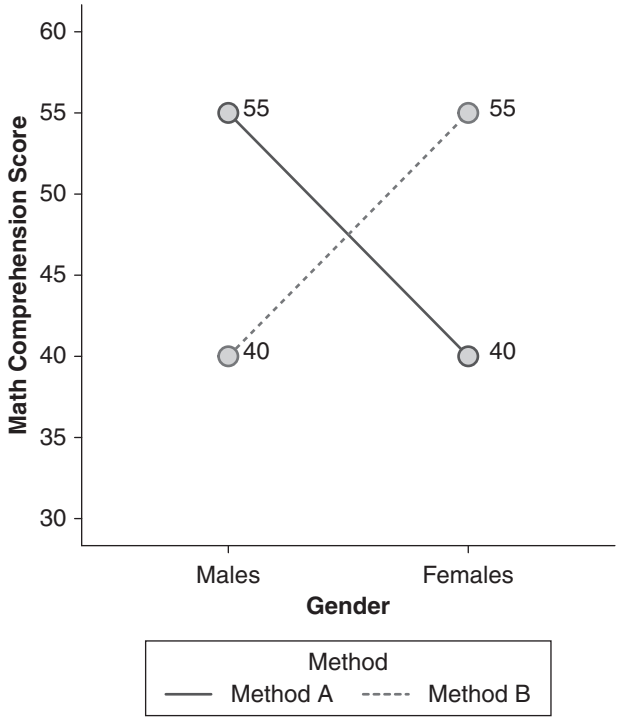



**Figure 3** Plot of Significant, Disordinal Interaction of Method by Gender

## Follow-Up Analyses for Significant Results

Significant main effects are often followed by post hoc tests to determine which pair or pairs of group means are significantly different. A wide variety of tests are available. They differ in their degree of adjustment for the compounding of Type I error (rejection of a true null hypothesis). Examples of a few post hoc tests are Fisher's LSD, Duncan, Newman–Keuls, Tukey, and Scheffé, in order from liberal (adjusts less, rejects more) to conservative (adjusts more, rejects less). Many other tests are available for circumstances when the group sample sizes are unequal or when group variances are unequal. For example, the Games Howell post hoc test is appropriate when the homogeneity of variance assumption is violated. Non-pairwise tests of means, sometimes called contrast analysis, can also be conducted. For example, a researcher might be interested in comparing a control group mean to a mean that represents the average of all other experimental groups combined.

When a significant interaction occurs in a study, main effects are not usually interpreted. When the effect of one factor is not constant across the levels of another factor, it is difficult to make generalized statements about a main effect. There are two categories of follow-up analysis for significant interactions. One is called simple main effects. It involves examining the effect of one factor at only one level of another factor. Suppose there are five treatment conditions in an experimental factor and two categories of age. Comparing the means for younger participants versus older participants within each condition level is one example of a series of tests for a simple main effect.

Another category of follow-up analysis is called interaction comparisons. This series of tests examines whether the difference in means between two levels of Factor A is equal across two levels of Factor B. Using the example in the above paragraph, one test in the series would involve comparing the difference in Condition A means for younger and older participants to the difference in Condition B

means for younger and older participants. Additional tests would compare the mean age differences for Conditions A and B, Conditions A and C, and so on.

### Effect Sizes

A final note about factorial designs concerns the practical significance of the results. Every research study involving factorial analysis of variance should include a measure of effect size. Tables are available for Cohen's *f* effect size. Descriptive labels have been attached to the *f* values (.1 is small, .25 is medium, and greater than .40 is large), although the magnitude of the effect size obtained in a study should be interpreted relative to other research in the particular field of study. Some widely available software programs report partial eta-squared values, but they overestimate actual effect sizes. Because of this positive bias, some researchers prefer to calculate omega-squared effect sizes.

*Carol S. Parke*

***See also*** Dependent Variable; Effect Size, Measures of; Independent Variable; Main Effects; Post Hoc Analysis; Repeated Measures Design

#### Further Readings

Glass, G. V, & Hopkins, K. D. (1996). *Statistical methods in education and psychology* (3rd ed.). Boston: Allyn & Bacon.

Green, S. B., & Salkind, N. J. (2005). *Using SPSS for Windows: Analyzing and understanding data* (4th ed.). Upper Saddle River, NJ: Prentice Hall. [see examples of conducting simple main effects and interaction comparisons for significant two-factor interactions]

Howell, D. C. (2002). *Statistical methods for psychology* (5th ed.). Pacific Grove, CA: Wadsworth.

Stevens, J. (2007). *Intermediate statistics: A modern approach* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Warner, R. M. (2008). *Applied statistics: From bivariate through multivariate techniques*. Thousand Oaks, CA: Sage.

## Factor Loadings

Factor loadings are part of the outcome from factor analysis, which serves as a data reduction method designed to explain the correlations between observed variables using a smaller number of factors. Because factor analysis is a widely used method in social and behavioral research, an in-depth examination of factor loadings and the related factor-loading matrix will facilitate a better understanding and use of the technique.

### Factor Analysis and Factor Loadings

Factor loadings are coefficients found in either a factor pattern matrix or a factor structure matrix. The former matrix consists of regression coefficients that multiply common factors to predict observed variables, also known as manifest variables, whereas the latter matrix is made up of product-moment correlation coefficients between common factors and observed variables.

The pattern matrix and the structure matrix are identical in orthogonal factor analysis where common factors are uncorrelated. This entry primarily examines factor loadings in this modeling situation, which is most commonly seen in applied research. Therefore, the majority of the entry content is devoted to factor loadings, which are both regression coefficients in the pattern matrix and correlation coefficients in the structure matrix. Factor loadings in oblique factor analysis are briefly discussed at the end of the entry, where common factors are correlated and the two matrices differ.

Besides, factor analysis could be exploratory (EFA) or confirmatory (CFA). EFA does not assume any model a priori, whereas CFA is designed to confirm a theoretically established factor model. Factor loadings play similar roles in these two modeling situations. Therefore, in this entry on factor loadings, the term *factor analysis* refers to both EFA and CFA, unless stated otherwise.

#### *Overview*

Factor analysis, primarily EFA, assumes that common factors do exist that are indirectly measured by observed variables, and that each observed variable is a weighted sum of common factors plus a unique component. Common factors are latent and they influence one or more observed variables. The unique component represents all those independent things, both systematic and random, that are specific to a particular observed

variable. In other words, a common factor is loaded by at least one observed variable, whereas each unique component corresponds to one and only one observed variable. Factor loadings are correlation coefficients between observed variables and latent common factors.

Factor loadings can also be viewed as standardized regression coefficients, or regression weights. Because an observed variable is a linear combination of latent common factors plus a unique component, such a structure is analogous to a multiple linear regression model where each observed variable is a response and common factors are predictors. From this perspective, factor loadings are viewed as standardized regression coefficients when all observed variables and common factors are standardized to have unit variance. Stated differently, factor loadings can be thought of as an optimal set of regression weights that predicts an observed variable using latent common factors.

Factor loadings usually take the form of a matrix, and this matrix is a standard output of almost all statistical software packages when factor analysis is performed. The factor loading matrix is usually denoted by the capital Greek letter $\Lambda$, or lambda, whereas its matrix entries, or factor loadings, are denoted by $\lambda_{ij}$ with $i$ being the row number and $j$ the column number. The number of rows of the matrix equals that of observed variables and the number of columns that of common factors.

### *Factor Loadings in a Hypothetical Example*

A typical example involving factor analysis is to use personality questionnaires to measure underlying psychological constructs. Item scores are observed data, and common factors correspond to latent personality attributes.

Suppose a psychologist is developing a theory that hypothesizes there are two personality attributes that are of interest, introversion and extroversion. To measure these two latent constructs, the psychologist develops a 5-item personality instrument and administers it to a randomly selected sample of 1,000 participants.

Thus, each participant is measured on five variables, and each variable can be modeled as a linear combination of the two latent factors plus a unique component. Stated differently, the score on an item for one participant, say, Participant A, consists of two parts. Part 1 is the average score on this item for all participants having identical levels of introversion and extroversion as Participant A, and this average item score is denoted by a constant times this participant's level of introversion plus a second constant times his or her level of extroversion. Part 2 is the unique component that indicates the amount of difference between the item score from Participant A and the said average item score. Obviously, such a two-part scenario is highly similar to a description of regression analysis.

In the above example, factor loadings for this item, or this observed variable, are nothing but the two constants that are used to multiply introversion and extroversion. There is a set of factor loadings for each item or each observed variable.

### *Factor Loadings in a Mathematical Form*

The mathematical form of a factor analysis model takes the following form:

$$\mathbf{x} = \mathbf{\Lambda f} + \boldsymbol{\varepsilon},$$

where $\mathbf{x}$ is a $p$-variate vector of standardized, observed data; $\mathbf{\Lambda}$ is a $p \times m$ matrix of factor loadings; $\mathbf{f}$ is an $m$-variate vector of standardized, common factors; and $\boldsymbol{\varepsilon}$ is a $p$-variate vector of standardized, unique components.

Back to the previous example, in which $p = 5$ and $m = 2$. So the factor loading matrix $\mathbf{\Lambda}$ is a $5 \times 2$ matrix consisting of correlation coefficients. The above factor analysis model can be written in another form for each item:

$$\begin{cases} x_1 = \lambda_{11} f_1 + \lambda_{12} f_2 + \varepsilon_1, \\ x_2 = \lambda_{21} f_1 + \lambda_{22} f_2 + \varepsilon_2, \\ x_3 = \lambda_{31} f_1 + \lambda_{32} f_2 + \varepsilon_3, \\ x_4 = \lambda_{41} f_1 + \lambda_{42} f_2 + \varepsilon_4, \\ x_5 = \lambda_{51} f_1 + \lambda_{52} f_2 + \varepsilon_5. \end{cases}$$

Each of the five equations corresponds to one item. In other words, each observed variable is represented by a weighted linear combination of common factors plus a unique component. And for each item, two factor loading constants bridge observed data and common factors. These constants are standardized regression weights because observed data, common factors, and unique components are all standardized to have zero mean

and unit variance. For example, in determining standardized $x_1$, $f_1$ is given the weight $\lambda_{11}$ and $f_2$ is given the weight $\lambda_{12}$, whereas in determining standardized $x_2$, $f_1$ is given the weight $\lambda_{21}$ and $f_2$ is given the weight $\lambda_{22}$.

The factor loading matrix can be used to define an alternative form of factor analysis model. Suppose the observed correlation matrix and the factor model correlation matrix are $\mathbf{R_X}$ and $\mathbf{R_F}$, respectively. The following alternative factor model can be defined using the factor loading matrix:

$$\mathbf{R_F} = \mathbf{\Lambda}\mathbf{\Lambda}^{\mathrm{T}} + \mathbf{\psi},$$

where $\mathbf{\psi}$ is a diagonal matrix of unique variances. Some factor analysis algorithms iteratively solve for $\mathbf{\Lambda}$ and $\mathbf{\psi}$ so that the difference between $\mathbf{R_X}$ and $\mathbf{R_F}$ is minimized.

### *Communality and Unique Variance*

Based on factor loadings, communality and unique variance can be defined. These two concepts relate to each observed variable.

The communality for an observed variable refers to the amount of variance in that variable that is explained by common factors. If the communality value is high, at least one of the common factors has a substantial impact on the observed variable. The sum of squared factor loadings is the communality value for that observed variable. And most statisticians use $h_i^2$ to denote the communality value for the $i$th observed variable.

The unique variance for an observed variable is computed as 1 minus that variable's communality value. The unique variance represents the amount of variance in that variable that is not explained by common factors.

## Issues Regarding Factor Loadings

### *Significance of Factor Loadings*

There are usually three approaches to the determination of whether or not a factor loading is significant: cutoff value, $t$ test, and confidence interval. However, it should be noted that the latter two are not commonly seen in applied research.

A factor loading that falls outside of the interval bounded by ($\pm$cutoff value) is considered to be large and is thus retained. On the other hand, a factor loading that does not meet the criterion indicates that the corresponding observed variable should not load on the corresponding common factor. The cutoff value is arbitrarily selected depending on the field of study, but ($\pm 0.4$) seems to be preferred by many researchers.

A factor loading can also be $t$ tested, and the null hypothesis for this test is that the loading is not significantly different from zero. The computed $t$ statistic is compared with the threshold chosen for statistical significance. If the computed value is larger than the threshold, the null is rejected in favor of the alternative hypothesis, which states that the factor loading differs significantly from zero.

A confidence interval (CI) can be constructed for a factor loading, too. If the CI does not cover zero, the corresponding factor loading is significantly different from zero. If the CI does cover zero, no conclusion can be made regarding the significance status of the factor loading.

### *Rotated Factor Loadings*

The need to rotate a factor solution relates to the factorial complexity of an observed variable, which refers to the number of common factors that have a significant loading for this variable. In applied research, it is desirable for an observed variable to load significantly on one and only one common factor, which is known as a simple structure. For example, a psychologist prefers to be able to place a questionnaire item into one and only one subscale.

When an observed variable loads on two or more factors, a factor rotation is usually performed to achieve a simple structure, which is a common practice in EFA. Of all rotation techniques, varimax is most commonly used. Applied researchers usually count on rotated factor loadings to interpret the meaning of each common factor.

### *Labeling Common Factors*

In EFA, applied researchers usually use factor loadings to label common factors. EFA assumes that common factors exist, and efforts are made to

determine the number of common factors and the set of observed variables that load significantly on each factor. The underlying nature of each such set of observed variables is used to give the corresponding common factor a name. For example, if a set of questionnaire items loads highly on a factor and the items refer to different aspects of extroversion, the common factor should be named accordingly to reflect the common thread that binds all of those items. However, the process of labeling common factors is very subjective.

### *Factor Loadings in Oblique Factor Analysis*

Oblique factor analysis is needed given the correlation of common factors. Unlike the orthogonal case, the factor pattern matrix and the factor structure matrix differ in this modeling situation. An examination of factor loadings involves interpreting both matrices in a combined manner. The pattern matrix provides information regarding the group of observed variables used to measure each common factor, thus contributing to an interpretation of common factors, whereas the structure matrix presents product-moment correlation coefficients between observed variables and common factors.

### *Factor Loadings in Second-Order Factor Analysis*

Factor-analyzing observed variables leads to a reduced number of first-order common factors, and the correlations between them can sometimes be explained further by an even smaller number of second-order common factors; and, a second-order factor model is usually analyzed under the CFA context. For this type of factor model, factor loadings refer to all of those regression/correlation coefficients that not only connect observed variables with first-order factors but also bridge two different levels of common factors.

*Hongwei Yang*

***See also*** Confirmatory Factor Analysis; Correlation; Exploratory Factor Analysis; Regression Coefficient; Structural Equation Modeling

**Further Readings**

Child, D. (1990). *The essentials of factor analysis* (2nd ed.). London: Cassel Educational Ltd.

Everitt, B. S. (1984). *An introduction to latent variable models*. London: Chapman & Hall.

Garson, G. D. (2008, March). Factor analysis. Retrieved July 26, 2008, from http://www2.chass.ncsu.edu/garson/pa765/factor.htm

Harman, H. H. (1976). *Modern factor analysis* (3rd ed.). Chicago: University of Chicago Press.

Lawley, D. N. (1940). The estimation of factor loadings by the method of maximum likelihood. *Proceedings of the Royal Society of Edinburgh, A-60,* 64–82.

Suhr, D. D. (n.d.). *Exploratory or confirmatory factor analysis*. Retrieved July 26, 2008, from http://www2.sas.com/proceedings/sugi31/200–31.pdf

## False Positive

The term *false positive* is most commonly employed in diagnostic classification within the context of assessing test validity. The term represents a diagnostic decision in which an individual has been identified as having a specific condition (such as an illness), when, in fact, he or she does not have the condition. The term *false positive* is less commonly used within the context of hypothesis testing to represent a Type I error, which is defined as rejection of a true null hypothesis, and thereby incorrectly concluding that the alternative hypothesis is supported. This entry focuses on the more common use of the term *false positive* within the context of diagnostic decision making. The disciplines that are most likely to be concerned with the occurrence of false positives are medicine, clinical psychology, educational and school psychology, forensic psychology (and the legal system), and industrial psychology. In each of the aforementioned disciplines, critical decisions are made about human beings based on the results of diagnostic tests or other means of assessment. The consequences associated with false positives can range from a person being found guilty of a murder he or she did not commit to a much less serious consequence, such as a qualified person being erroneously identified as unsuitable for a job, and thus not being offered the job.

## Basic Definitions

Within the framework of developing measuring instruments that are capable of categorizing people and/or predicting behavior, researchers attempt to optimize correct categorizations (or predictions) and minimize incorrect categorizations (or predictions). Within the latter context, the following four diagnostic decisions are possible (the first two of which are correct and the latter two incorrect): *true positive, true negative, false positive, false negative*. The terms *true* and *false* in each of the aforementioned categories designate whether or not a diagnostic decision made with respect to an individual is, in fact, correct (as in the case of a true positive and a true negative) or incorrect (as in the case of a false positive and a false negative). The terms *positive* and *negative* in each of the aforementioned categories refer to whether or not the test result obtained for an individual indicates he or she has the condition in question. Thus, both a true positive and false positive represent individuals who obtain a positive test result—the latter indicating that such individuals have the condition in question. On the other hand, both a true negative and a false negative represent individuals who obtain a negative test result—the latter indicating that such individuals do not have the condition in question.

In the discipline of medicine, the true positive rate for a diagnostic test is referred to as the *sensitivity* of the test—that is, the probability that a person will test positive for a disease, given the person actually has the disease. The true negative rate for a diagnostic test is referred to as the *specificity* of the test—that is, the probability that a person will test negative for a disease, given the person actually does not have the disease. As a general rule, in order for a diagnostic test to be a good instrument for detecting the presence of a disease, it should be high in both sensitivity and specificity. The proportion of true positives and false positives in a population is referred to as the *selection ratio* because it represents the proportion of the population that is identified as possessing the condition in question.

It was noted earlier that a false positive can also be employed within the context of hypothesis testing to represent a Type I error. Analogously, a false negative can be employed within the context of hypothesis testing to represent a Type II error—the latter being failure to reject a false null hypothesis and thereby concluding incorrectly that the alternative hypothesis is not supported. A true negative can be employed within the context of hypothesis testing to represent retention of a correct null hypothesis, whereas a true positive can be employed to represent rejection of a false null hypothesis. Recollect that within the context of hypothesis testing, a null hypothesis states that no experimental effect is present, whereas the alternative hypothesis states that an experimental effect is present. Thus, retention of a null hypothesis is analogous to a clinical situation in which it is concluded a person is normal, whereas rejection of the null hypothesis is analogous to reaching the conclusion a person is not normal.

## Illustrative Examples

Three commonly encountered examples involving testing will be used to illustrate the four diagnostic decisions. The first example comes from the field of medicine, where diagnostics tests are commonly employed in making decisions regarding patients. Thus, it will be assumed that the condition a diagnostic test is employed to identify is a physical or psychological illness. In such a case, a *true positive* is a person whom the test indicates has the illness and does, in fact, have the illness. A *true negative* is a person whom the test indicates does not have the illness and, in fact, does not have the illness. A *false positive* is a person whom the test indicates has the illness but, in fact, does not have the illness. A *false negative* is a person whom the test indicates does not have the illness but, in fact, has the illness.

The second example involves the use of the polygraph for the purpose of ascertaining whether a person is responding honestly. Although in most states, polygraph evidence is not generally admissible in court, a person's performance on a polygraph can influence police and prosecutors with regard to their belief concerning the guilt or innocence of an individual. The condition that the polygraph is employed to identify is whether or not a person is responding honestly to what are considered to be relevant questions. In the case of a polygraph examination, a *true positive* is a person whom the polygraph identifies as dishonest and is, in fact, dishonest. A *true negative* is a person whom the

polygraph identifies as honest and is, in fact, honest. A *false positive* is a person whom the polygraph identifies as dishonest but is, in fact, honest. A *false negative* is a person whom the polygraph identifies as honest but is, in fact, dishonest.

The final example involves the use of an integrity test, which is commonly used in business and industry in assessing the suitability of a candidate for a job. Although some people believe that integrity tests are able to identify individuals who will steal from an employer, the more general consensus is that such tests are more likely to identify individuals who will not be conscientious employees. In the case of an integrity test, the condition that the test is employed to identify is the unsuitability of a job candidate. With regard to a person's performance on an integrity test, a *true positive* is a person whom the test identifies as an unsuitable employee and, in fact, will be an unsuitable employee. A *true negative* is a person whom the test identifies as a suitable employee and, in fact, will be a suitable employee. A *false positive* is a person whom the test identifies as an unsuitable employee but, in fact, will be a suitable employee. A *false negative* is a person whom the test identifies as a suitable employee but, in fact, will be an unsuitable employee.

## Relative Seriousness of False Positive Versus False Negative

It is often the case that the determination of a cutoff score on a test (or criterion of performance on a polygraph) for deciding to which category a person will be assigned will be a function of the perceived seriousness of incorrectly categorizing a person a false positive versus a false negative. Although it is not possible to state that one type of error will always be more serious than the other, a number of observations can be made regarding the seriousness of the two types of errors. The criterion for determining the seriousness of an error will always be a function of the consequences associated with the error. In medicine, physicians tend to view a false negative as a more serious error than a false positive the latter being consistent with the philosophy that it is better to treat a nonexistent illness than to neglect to treat a potentially serious illness. Yet things are not always that clear-cut. As an example, although the consequence of failure to diagnose breast cancer (a false negative) could cost a woman her life, the consequences associated with a woman being a false positive could range from minimal (e.g., the woman is administered a relatively benign form of chemotherapy) to severe (e.g., the woman has an unnecessary mastectomy).

In contrast to medicine, the American legal system tends to view a false positive as a more serious error than a false negative. The latter is reflected in the use of the "beyond a reasonable doubt" standard in criminal courts, which reflects the belief that it is far more serious to find an innocent person guilty than to find a guilty person innocent. Once again, however, the consequences associated with the relative seriousness of both types of errors may vary considerably depending upon the nature of the crime involved. For example, one could argue that finding a serial killer innocent (a false negative) constitutes a far more serious error than wrongly convicting an innocent person of a minor felony (a false positive) that results in a suspended sentence.

## The Low Base Rate Problem

The *base rate* of a behavior or medical condition is the frequency with which it occurs in a population. The *low base rate problem* occurs when a diagnostic test that is employed to identify a low base rate behavior or condition tends to yield a disproportionately large number of false positives. Thus, when a diagnostic test is employed in medicine to detect a rare disease, it may, in fact, identify virtually all of the people who are afflicted with the disease, but in the process erroneously identify a disproportionately large number of healthy people as having the disease, and because of the latter, the majority of people labeled positive will, in fact, not have the disease.

The relevance of the low base rate problem to polygraph and integrity testing is that such instruments may correctly identify most guilty individuals and potentially unsuitable employees, yet, in the process, erroneously identify a large number of innocent people as guilty or, in the case of an integrity test, a large number of potentially suitable employees as unsuitable. Estimates of false positive rates associated with the polygraph and integrity tests vary substantially, but critics of the

latter instruments argue that error rates are unacceptably high. In the case of integrity tests, people who utilize them may concede that although such tests may yield a large number of false positives, at the same time they have a relatively low rate of false negatives. Because of the latter, companies that administer integrity tests cite empirical evidence that such tests are associated with a decrease in employee theft and an increase in productivity. In view of the latter, they consider the consequences associated with a false positive (not hiring a suitable person) to be far less damaging to the company than the consequences associated with a false negative (hiring an unsuitable person).

### Use of Bayes's Theorem for Computing a False Positive Rate

In instances where the false positive rate for a test cannot be determined from empirical data, Bayes's theorem can be employed to estimate the latter. Bayes's theorem is a rule for computing conditional probabilities that was stated by an 18th-century English clergyman, the Reverend Thomas Bayes. A conditional probability is the probability of Event A given the fact that Event B has already occurred. Bayes's theorem assumes there are two sets of events. In the first set, there are $n$ events to be identified as $A_1, A_2, \ldots, A_n$, and in the second set, there are two events to be identified as $B+$ and $B-$. Bayes's theorem allows for the computation of the probability that $A_j$ (where $1 \leq j \leq n$) will occur, given it is known that $B+$ has occurred. As an example, the conditional probability $P(A_2/B+)$ represents the probability that Event $A_2$ will occur, given the fact that Event $B+$ has already occurred.

An equation illustrating the application of Bayes's theorem is presented below. In the latter equation, it is assumed that Set 1 is comprised of the two events $A_1$ and $A_2$ and Set 2 is comprised of the two events $B+$ and $B-$. If it is assumed $A_1$ represents a person who is, in fact, sick; $A_2$ represents a person who is, in fact, healthy; $B+$ indicates a person who received a positive diagnostic test result for the illness in question; and B indicates a person who received a negative diagnostic test result for the illness in question, then the conditional probability $P(A_2/B+)$ computed with Bayes's theorem represents the probability that a person will be healthy given that his or her diagnostic test result was positive.

$$P(A_2/B+) = \frac{P(B+/A_2)P(A_2)}{P(B+/A_1)P(A_1) + P(B+/A_2)P(A_2)}$$

When the above noted conditional probability $P(A_2/B+)$ is multiplied by $P(B+)$ (the proportion of individuals in the population who obtain a positive result on the diagnostic test), the resulting value represents the proportion of false positives in the population. In order to compute $P(A_2/B+)$, it is necessary to know the population base rates $A_1$ and $A_2$ as well as the conditional probabilities $P(B+/A_1)$ and $P(B+/A_2)$. Obviously, if one or more of the aforementioned probabilities is not known or cannot be estimated accurately, computing a false positive rate will be problematical.

*David J. Sheskin*

***See also*** Bayes's Theorem; Sensitivity; Specificity; True Positive

#### Further Readings

Feinstein, A. R. (2002). *Principles of medical statistics*. Boca Raton, FL: Chapman & Hall/CRC.

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions* (3rd ed.). Hoboken, NJ: Wiley-Interscience.

Meehl, P., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin, 52,* 194 216.

Pagano, M., & Gauvreau, K. (2000). *Principles of biostatistics* (2nd ed.). Pacific Grove, CA: Duxbury.

Rosner, B. (2006). *Fundamentals of biostatistics* (6th ed.). Belmont, CA: Thomson-Brooks/Cole.

Sheskin, D. J. (2007). *Handbook of parametric and nonparametric statistical procedures* (4th ed.). Boca Raton, FL: Chapman & Hall/CRC.

Wiggins, J. (1973). *Personality and prediction: Principles of personality assessment*. Menlo Park, CA: Addison Wesley.

## FALSIFIABILITY

The concept of falsifiability is central to distinguishing between systems of knowledge and

understanding, specifically between scientific theories of understanding the world and those considered nonscientific. The importance of the concept of falsifiability was developed most thoroughly by the philosopher Karl Popper in the treatise *Conjectures and Refutations: The Growth of Scientific Knowledge*. Specifically, falsifiability refers to the notion that a theory or statement can be found to be false; for instance, as the result of an empirical test.

Popper sought to distinguish between various means of understanding the world in an effort to determine what constitutes a scientific approach. Prior to his seminal work, merely the empirical nature of scientific investigation was accepted as the criterion that differentiated it from pseudo- or nonscientific research. Popper's observation that many types of research considered nonscientific were also based upon empirical techniques led to dissatisfaction with this conventional explanation. Consequently, several empirically based methods colloquially considered scientific were contrasted in an effort to determine what distinguished science from pseudoscience. Examples chosen by Popper to illustrate the diversity of empirical approaches included physics, astrology, Marxian theories of history, and metaphysical analyses. Each of these epistemic approaches represents a meaningful system of interpreting and understanding the world around us, and has been used earnestly throughout history with varying degrees of perceived validity and success.

Popper used the term *line of demarcation* to distinguish the characteristics of scientific from nonscientific (pseudoscientific) systems of understanding. What Popper reasoned differentiated the two categories of understanding is that the former could be falsified (or found to be *not* universally true), whereas the latter was either incapable of being falsified or had been used in such a way that renders falsification unlikely. According to Popper, this usage takes the form of seeking corroboratory evidence to verify the verisimilitude of a particular pseudoscientific theory. For example, with respect to astrology, proponents subjectively interpret events (data) in ways that corroborate their preconceived astrological theories and predictions, rather than attempting to find data that undermine the legitimacy of astrology as an epistemic enterprise.

Popper found similarity between astrologists and those who interpret and make predictions about historical events via Marxian analyses in that both have historically sought to verify rather than falsify their perspectives as a matter of practice. Where a lack of corroboration between reality and theory exists, proponents of both systems reinterpret their theoretical position so as to correspond with empirical observations, essentially undermining the extent to which the theoretical perspective can be falsified. The proponents of both pseudoscientific approaches tacitly accept the manifest truth of their epistemic orientations irrespective of the fact that apparent verisimilitude is contingent upon subjective interpretations of historical events.

Popper rejected the notion that scientific theories were those thought most universally true, given the notion that verifying theories in terms of their correspondence to the truth is a quixotic task requiring omniscience. According to Popper, one cannot predict the extent to which future findings could falsify a theory, and searching for verification of the truth of a given theory ignores this potentiality. Instead of locating the essence of science within a correspondence with truth, Popper found that theories most scientific were those capable of being falsified. This renders all scientific theories tenable at best, in the sense that the most plausible scientific theories are merely those that have *yet* to be falsified.

Every empirical test of a theory is an attempt to falsify it, and there are degrees of testability with respect to theories as a whole. Focusing on falsification relocates power from the extent to which a theory corresponds with a given reality or set of circumstances to the extent to which it logically can be proven false given an infinite range of empirical possibilities. Contrarily, a hypothetical theory that is capable of perfectly and completely explaining a given phenomenon is inherently unscientific because it cannot be falsified logically. Where theories are reinterpreted to make them more compatible with potentially falsifying empirical information, it is done to the benefit of its correspondence with the data, but to the detriment of the original theory's claim to scientific status.

As an addendum, Popper rejected the notion that only tenable theories are most useful, because those that have been falsified may illuminate

constructive directions for subsequent research. Thus, the principle of falsificationism does not undermine the inherent meaning behind statements that fall short of achieving its standard of scientific status.

Some competing lines of demarcation in distinguishing scientific from pseudoscientific research include the verificationist and anarchistic epistemological perspectives. As previously noted, the simple standard imposed by verificationism states that a theory is considered scientific merely if it can be verified through the use of empirical evidence. A competing line involves Paul Feyerabend's anarchistic epistemological perspective, which holds that any and all statements and theories can be considered scientific because history shows that "whatever works" has been labeled scientific regardless of any additional distinguishing criteria.

*Douglas J. Dallier*

***See also*** External Validity; Hypothesis; Internal Validity; *Logic of Scientific Discovery, The*; Research Design Principles; Test; Theory

**Further Readings**

Feyerabend, P. (1975). *Against method: Outline of an anarchistic theory of knowledge*. London: New Left Books.

Kuhn, T. (1962). *The structure of scientific revolutions*. Chicago: University of Chicago Press.

Lakatos, I., Feyerabend, P., & Motterlini, M. (1999). *For and against method: Including Lakatos's lectures on scientific method and the Lakatos-Feyerabend correspondence*. Chicago: University of Chicago Press.

Mace, C. A. (Ed.). (1957). *Philosophy of science: A personal report: British philosophy in mid-century*. London: Allen and Unwin.

Popper, K. (1962). *Conjectures and refutations: The growth of scientific knowledge*. New York: Basic Books.

# Field Study

A field study refers to research that is undertaken in the real world, where the confines of a laboratory setting are abandoned in favor of a natural setting. This form of research generally prohibits the direct manipulation of the environment by the researcher. However, sometimes, independent and dependent variables already exist within the social structure under study, and inferences can then be drawn about behaviors, social attitudes, values, and beliefs. It must be noted that a field study is separate from the concept of a field experiment. Overall, field studies belong to the category of nonexperimental designs where the researcher uses what already exists in the environment. Alternatively, field experiments refer to the category of experimental designs where the researcher follows the scientific process of formulating and testing hypotheses by invariably manipulating some aspect of the environment. It is important that prospective researchers understand the types, aims, and issues; the factors that need to be considered; and the advantages and concerns raised when conducting the field study type of research.

Field studies belong to the category of nonexperimental design. These studies include the *case study*—an in-depth observation of one organization, individual, or animal; *naturalistic observation*—observation of an environment without any attempt to interfere with variables; *participant observer study*—observation through the researcher's submergence into the group under study; and *phenomenology*—observation derived from the researcher's personal experiences. The two specific aims of field studies are exploratory research and hypothesis testing. Exploratory research seeks to examine what exists in order to have a better idea about the dynamics that operate within the natural setting. Here, the acquisition of knowledge is the main objective. With hypothesis testing, the field study seeks to determine whether the null hypothesis or the alternative hypothesis best predicts the relationship of variables in the specific context; assumptions can then be used to inform future research.

## Real-Life Research and Applications

Field studies have often provided information and reference points that otherwise may not have been available to researchers. For example, the famous obedience laboratory experiment by Stanley Milgram was criticized on the grounds that persons in real-life situations would not unquestioningly carry out unusual requests by persons perceived to

be authority figures as they did in the laboratory experiment. Leonard Bickman then decided to test the obedience hypothesis using a real-life application. He found that his participants were indeed more willing to obey the stooge who was dressed as a guard than the one who dressed as a sportsman or a milkman. Another example of field research usage is Robert Cialidini's investigation of how some professionals, such as con men, sales representatives, politicians, and the like, are able to gain compliance from others. In reality, he worked in such professions and observed the methods that these persons used to gain compliance from others. From his actual experiences, he was able to offer six principles that cover the compliance techniques used by others. Some field studies take place in the workplace to test attitudes and efficiency. Therefore, field studies can be conducted to examine a multitude of issues that include playground attitudes of children, gang behaviors, how people respond to disasters, efficiency of organization protocol, and even behavior of animals in their natural environment. Information derived from field studies result in correlational interpretations.

## Strengths and Weaknesses

Field studies are employed in order to increase ecological and external validity. Because variables are not directly manipulated, the conclusions drawn are deemed to be true to life and generalizable. Also, such studies are conducted when there is absolutely no way of even creating mundane realism in the laboratory. For example, if there is a need to investigate looting behavior and the impact of persons on each other to propel this behavior, then a laboratory study cannot suffice for the investigation because of the complexity of the variables that may be involved. Field research is therefore necessary.

Although field studies are nonexperimental, this does not imply that such studies are not empirical. Scientific rigor is promoted by various means, including the methods of data collection used in the study. Data can be reliably obtained through direct observation, coding, note-taking, the use of interview questions—preferably structured—and audiovisual equipment to garner information. Even variables such as the independent variable, dependent variable, and other specific variables of interest that already operate in the natural setting may be identified and, to a lesser extent, controlled by the researcher because those variables would become the focus of the study. Overall, field studies tend to capture the essence of human behavior, particularly when the persons under observation are unaware that they are being observed, so that authentic behaviors are reflected without the influence of demand characteristics (reactivity) or social desirability answers. Furthermore, when observation is unobtrusive, the study's integrity is increased.

However, because field studies, by their very nature, do not control extraneous variables, it is exceedingly difficult to ascertain which factor or factors are more influential in any particular context. Bias can also be an issue if the researcher is testing a hypothesis. There is also the problem of replication. Any original field study sample will not be accurately reflective of any other replication of that sample. Furthermore, there is the issue of ethics. Many times, to avoid reactivity, researchers do not ask permission from their sample to observe them, and this may cause invasion-of-privacy issues even though such participants are in the public eye. For example, if research is being carried out about the types of kissing that take place in a park, even though the persons engaged in kissing are doing so in public, had they known that their actions were being videotaped, they may have strongly objected. Other problems associated with field studies include the fact that they can be quite time-consuming and expensive, especially if a number of researchers are required as well as audiovisual technology.

*Indeira Persaud*

***See also*** Ecological Validity; Nonexperimental Design; Reactive Arrangements

### Further Readings

Allen, M. J. (1995). *Introduction to psychological research*. Itasca, IL: F. E. Peacock.

Babbie, E. (2004). *The practice of social research* (10th ed.). Belmont, CA: Thomson Wadsworth.

Bickman, L. (1974). Clothes make the person. *Psychology Today, 8*(4), 48–51.

Cialdini, R. B. (2006). *Influence: The psychology of persuasion*. New York: HarperCollins.
Robson, C. (2003). *Real world research*. Oxford, UK: Blackwell.
Solso, R. L., Johnson, H. H., & Beal, M. K. (1998). *Experimental psychology: A case approach*. New York: Longman.

# File Drawer Problem

The file drawer problem is the threat that the empirical literature is biased because nonsignificant research results are not disseminated. The consequence of this problem is that the results available provide a biased portrayal of what is actually found, so literature reviews (including meta-analyses) will conclude stronger effects than actually exist. The term arose from the image that these nonsignificant results are placed in researchers' file drawers, never to be seen by others. This file drawer problem also has several similar names, including *publication* or *dissemination bias*. Although all literature reviews are vulnerable to this problem, meta-analysis provides methods of detecting and correcting for this bias. This entry first discusses the sources of publication bias and then the detection and correction of such bias.

## Sources

The first source of publication bias is that researchers may be less likely to submit null than significant results. This tendency may arise in several ways. Researchers engaging in "data snooping" (cursory data analyses to determine whether more complete pursuit is warranted) simply may not pursue investigation of null results. Even when complete analyses are conducted, researchers may be less motivated—due to expectations that the results will not be published, professional pride, or financial interest in finding supportive results—to submit results for publication.

The other source is that null results are less likely to be accepted for publication than are significant results. This tendency is partly due to reliance on decision making from a null hypothesis significance testing (versus effect size) framework; statistically significant results lead to conclusions, whereas null results are inconclusive. Reviewers who have a professional or financial interest in certain results may also be less accepting of and more critical toward null results than those that confirm their expectations.

## Detection

Three methods are commonly used to evaluate whether publication bias exists within a literature review. Although one of these methods can be performed using vote-counting approaches to research synthesis, these approaches are typically conducted within a meta-analysis focusing on effect sizes.

The first method is to compare results of published versus unpublished studies, if the reviewer has obtained at least some of the unpublished studies. In a vote-counting approach, the reviewer can evaluate whether a higher proportion of published studies finds a significant effect than do the proportion of unpublished studies. In a meta-analysis, one performs moderator analyses that statistically compare whether effect sizes are greater in published versus unpublished studies. An absence of differences is evidence against a file drawer problem.

A second approach is through the visual examination of funnel plots, which are scatterplots of each study's effect size to sample size. Greater variability of effect sizes is expected in smaller versus larger studies, given their greater sampling variability. Thus, funnel plots are expected to look like an isosceles triangle, with a symmetric distribution of effect sizes around the mean across all levels of sample size. However, small studies that happen to find small effects will not be able to conclude statistical significance and therefore may be less likely to be published. The resultant funnel plot will be asymmetric, with an absence of studies in the small sample size/small effect size corner of the triangle.

A third, related approach is to compute the correlation between effect sizes and sample sizes across studies. In the absence of publication bias, one expects no correlation; small and large studies should find similar effect sizes. However, if nonsignificant results are more likely relegated to the file drawer, then one would find that only the small studies finding large effects are published. This would result in a correlation between sample size and effect size (a negative correlation if the average

effect size is positive and a positive correlation if the average effect size is negative).

### Correction

There are four common ways of correcting for the file drawer problem. The first is not actually a correction, but an attempt to demonstrate that the results of a meta-analysis are robust to this problem. This approach involves computing a failsafe number, which represents the number of studies with an average effect size of zero that could be added to a meta-analysis before the average effect becomes nonsignificant. If the number is large, one concludes that it is not realistic that so many excluded studies could exist so as to invalidate the conclusions, so the review is robust to the file drawer problem.

A second approach is to exclude underpowered studies from a literature review. The rationale for this suggestion is that if the review includes only studies of a sample size large enough to detect a predefined effect size, then nonsignificant results should not result in publication bias among this defined set of studies. This suggestion assumes that statistical nonsignificance is the primary source of unpublished research. This approach has the disadvantage of excluding a potentially large number of studies with smaller sample sizes, and therefore might often be an inefficient solution.

A third way to correct for this problem is through trim-and-fill methods. Although several variants exist, the premise of these methods is a two-step process based on restoring symmetry to a funnel plot. First, one "trims" studies that are in the represented corner of the triangle until a symmetric distribution is obtained; the mean effect size is then computed from this subset of studies. Second, one restores the trimmed studies and "fills" the missing portion of the funnel plot by imputing studies to create symmetry; the heterogeneity of effect sizes is then estimated from this filled set.

A final method of management is through a family of selection (weighted distribution) models. These approaches use a distribution of publication likelihood at various levels of statistical significance to weight the observed distribution of effect sizes for publication bias. These models are statistically complex, and the field has not reached agreement on best practices in their use. One challenge is that the user typically must specify a selection model, often with little information.

*Noel A. Card*

***See also*** Effect Size, Measures of; Literature Review; Meta-Analysis

#### Further Readings

Begg, C. B. (1994). Publication bias. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 399–409). New York: Russell Sage Foundation.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin, 86,* 638–641.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.). (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. Hoboken, NJ: Wiley.

## Fisher's Least Significant Difference Test

When an analysis of variance (ANOVA) gives a significant result, this indicates that at least one group differs from the other groups. Yet the omnibus test does not indicate which group differs. In order to analyze the pattern of difference between means, the ANOVA is often followed by specific comparisons, and the most commonly used involves comparing two means (the so-called pairwise comparisons).

The first pairwise comparison technique was developed by Ronald Fisher in 1935 and is called the *least significant difference* (LSD) test. This technique can be used only if the ANOVA $F$ omnibus is significant. The main idea of the LSD is to compute the smallest significant difference (i.e., the LSD) between two means as if these means had been the only means to be compared (i.e., with a $t$ test) and to declare significant any difference larger than the LSD.

### Notations

The data to be analyzed comprise $A$ groups, and a given group is denoted $a$. The number of observations of the $a$th group is denoted $S_a$. If all groups have the same size, the notation $S$ is used. The

total number of observations is denoted $N$. The mean of Group $a$ is denoted $M_{a+}$. From the ANOVA, the mean square of error (i.e., within group) is denoted $MS_{S(A)}$ and the mean square of effect (i.e., between group) is denoted $MS_A$.

## Least Significant Difference

The rationale behind the LSD technique value comes from the observation that when the null hypothesis is true, the value of the $t$ statistics evaluating the difference between Groups $a$ and $a'$ is equal to

$$t = \frac{M_{a+} - M_{a'+}}{\sqrt{MS_{S(A)}\left(\frac{1}{S_a} + \frac{1}{S_{a'}}\right)}} \quad (1)$$

and follows a Student's $t$ distribution with $N - A$ degrees of freedom. The ratio $t$ therefore would be declared significant at a given $\alpha$ level if the value of $t$ is larger than the critical value for the $\alpha$ level obtained from the $t$ distribution and denoted $t_{\nu,\alpha}$ (where $\nu = N - A$ is the number of degrees of freedom of the error; this value can be obtained from a standard $t$ table). Rewriting this ratio shows that a difference between the means of Groups $a$ and $a'$ will be significant if

$$|M_{a+} - M_{a'+}| > \text{LSD} = t_{\nu,\alpha}\sqrt{MS_{S(A)}\left(\frac{1}{S_a} + \frac{1}{S_{a'}}\right)} \quad (2)$$

When there is an equal number of observations per group, Equation 2 can be simplified as

$$\text{LSD} = t_{\nu,\alpha}\sqrt{MS_{S(A)}\frac{2}{S}} \quad (3)$$

In order to evaluate the difference between the means of Groups $a$ and $a'$ (where $a$ and $a'$ are the indices of the two groups under consideration), we take the absolute value of the difference between the means and compare it to the value of LSD. If

$$|M_{i+} - M_{j+}| \geq \text{LSD} \quad (4)$$

then the comparison is declared significant at the chosen $\alpha$ level (usually .05 or .01). Then, this procedure is repeated for all $\frac{A(A-1)}{2}$ comparisons.

Note that LSD has more power compared to other post hoc comparison methods (e.g., the honestly significant difference test, or Tukey test) because the $\alpha$ level for each comparison is not corrected for multiple comparisons. And, because LSD does not correct for multiple comparisons, it severely inflates Type I error (i.e., finding a difference when it does not actually exist). As a consequence, a revised version of the LSD test has been proposed by Anthony J. Hayter (and is known as the *Fisher–Hayter* procedure) where the modified LSD (MLSD) is used instead of the LSD. The MLSD is computed using the Studentized range distribution $q$ as

$$\text{MLSD} = q_{\alpha,A-1}\sqrt{\frac{MS_{S(A)}}{S}} \quad (5)$$

where $q_{\alpha,A-1}$ is the $\alpha$-level critical value of the Studentized range distribution for a range of $A - 1$ and for $\nu = N - A$ degrees of freedom. The MLSD procedure is more conservative than the LSD, but more powerful than the Tukey approach because the critical value for the Tukey approach is obtained from a Studentized range distribution equal to $A$. This difference in range makes Tukey's critical value always larger than the one used for the MLSD, and therefore, it makes Tukey's approach more conservative.

## Example

In a series of experiments on eyewitness testimony, Elizabeth Loftus wanted to show that the wording of a question influenced witnesses' reports. She showed participants a film of a car accident, then asked them a series of questions. Among the questions was one of five versions of a critical question asking about the speed the vehicles were traveling:

1. How fast were the cars going when they *hit* each other?
2. How fast were the cars going when they *smashed into* each other?
3. How fast were the cars going when they *collided with* each other?
4. How fast were the cars going when they *bumped* each other?
5. How fast were the cars going when they *contacted* each other?

**Table 1** Results for a Fictitious Replication of Loftus and Palmer (1974) in Miles per Hour

| | *Contact* | *Hit* | *Bump* | *Collide* | *Smash* |
|---|---|---|---|---|---|
| | 21 | 23 | 35 | 44 | 39 |
| | 20 | 30 | 35 | 40 | 44 |
| | 26 | 34 | 52 | 33 | 51 |
| | 46 | 51 | 29 | 45 | 47 |
| | 35 | 20 | 54 | 45 | 50 |
| | 13 | 38 | 32 | 30 | 45 |
| | 41 | 34 | 30 | 46 | 39 |
| | 30 | 44 | 42 | 34 | 51 |
| | 42 | 41 | 50 | 49 | 39 |
| | 26 | 35 | 21 | 44 | 55 |
| *M.+* | *30* | *35* | *38* | *41* | *46* |

**Table 2** ANOVA Results for the Replication of Loftus and Palmer (1974).

| *Source* | *df* | *SS* | *MS* | *F* | *Pr(F)* |
|---|---|---|---|---|---|
| Between: *A* | 4 | 1,460 | 365 | 4.56 | .0036 |
| Error: *S*(*A*) | 45 | 3,600 | 80 | | |
| Total | 49 | 5,060 | | | |

The data from a fictitious replication of Loftus' experiment are shown in Table 1. We have $A = 4$ groups and $S = 10$ participants *per* group.

The ANOVA found an effect of the verb used on participants' responses. The ANOVA table is shown in Table 2.

### *Least Significant Difference*

For an $\alpha$ level of .05, the LSD for these data is computed as

$$\begin{aligned} \text{LSD} &= t_{\nu,\,.05}\sqrt{MS_{S(A)}\frac{2}{n}} \\ &= t_{\nu,\,.05}\sqrt{80.00 \times \frac{2}{10}} \\ &= 2.01\sqrt{\frac{160}{10}} \\ &= 2.01 \times 4 \\ &= 8.04 \end{aligned} \tag{6}$$

**Table 3** LSD: Differences Between Means and Significance of Pairwise Comparisons From the (Fictitious) Replication of Loftus and Palmer (1974)

| | *Experimental Group* | | | | |
|---|---|---|---|---|---|
| | $M_{1.+}$ = *Contact* | $M_{2.+}$ = *Hit* | $M_{3.+}$ = *Bump* | $M_{4.+}$ = *Collide* | $M_{5.+}$ = *Smash* |
| $M_{1.+} = 30$ Contact | 0.00 | 5.00 *ns* | 8.00 *ns* | 11.00** | 16.00** |
| $M_{2.+} = 35$ Hit | | 0.00 | 3.00 *ns* | 6.00 *ns* | 11.00** |
| $M_{3.+} = 38$ Bump | | | 0.00 | 3.00 *ns* | 8.00 *ns* |
| $M_{4.+} = 41$ Collide | | | | 0.00 | 5.00 *ns* |
| $M_{5.+} = 46$ Smash | | | | | 0.00 |

*Notes:* Differences larger than 8.04 are significant at the $\alpha = .05$ level and are indicated with *, and differences larger than 10.76 are significant at the $\alpha = .01$ level and are indicated with **.

**Table 4** MLSD: Differences Between Means and Significance of Pairwise Comparisons From the (Fictitious) Replication of Loftus and Palmer (1974)

| | *Experimental Group* | | | | |
|---|---|---|---|---|---|
| | $M_{1.+}$ = *Contact* | $M_{2.+}$ = *Hit* | $M_{3.+}$ = *Bump* | $M_{4.+}$ = *Collide* | $M_{5.+}$ = *Smash* |
| $M_{1.+} = 30$ Contact | 0.00 | 5.00 *ns* | 8.00 *ns* | 11.00** | 16.00** |
| $M_{2.+} = 35$ Hit | | 0.00 | 3.00 *ns* | 6.00 *ns* | 11.00** |
| $M_{3.+} = 38$ Bump | | | 0.00 | 3.00 *ns* | 8.00 *ns* |
| $M_{4.+} = 41$ Collide | | | | 0.00 | 5.00 *ns* |
| $M_{5.+} = 46$ Smash | | | | | 0.00 |

*Notes:* Differences larger than 10.66 are significant at the $\alpha = .05$ level and are indicated with *, and differences larger than 13.21 are significant at the $\alpha = .01$ level and are indicated with **.

A similar computation will show that, for these data, the LSD for an $\alpha$ level of .01 is equal to LSD $= 2.69 \times 4 = 10.76$.

For example, the difference between $M_{\text{contact}+}$ and $M_{\text{hit}+}$ is declared nonsignificant because

$$|M_{\text{contact}+} - M_{\text{hit}+}| = |30 - 35| = 5 < 8.04. \quad (7)$$

The differences and significance of all pairwise comparisons are shown in Table 3.

### *Modified Least Significant Difference*

For an $\alpha$ level of .05, the value of $q_{.05,A-1}$ is equal to 3.77 and the MLSD for these data is computed as

$$MLSD = q_{\alpha,A-1}\sqrt{\frac{MS_{S(A)}}{S}} = 3.77 \times \sqrt{8} = 10.66 \quad (8)$$

The value of $q_{.01,A-1} = 4.67$, and a similar computation will show that, for these data, the MLSD for an $\alpha$ level of .01 is equal to $MLSD = 4.67 \times \sqrt{8} = 13.21$.

For example, the difference between $M_{\text{contact}+}$ and $M_{\text{hit}+}$ is declared nonsignificant because

$$|M_{\text{contact}+} - M_{\text{hit}+}| = |30 - 35| = 5 < 10.66. \quad (9)$$

The differences and significance of all pairwise comparisons are shown in Table 4.

*Lynne J. Williams and Hervé Abdi*

***See also*** Analysis of Variance (ANOVA); Bonferroni Procedure; Honestly Significant Difference (HSD) Test; Multiple Comparison Tests; Newman-Keuls Test and Tukey Test; Pairwise Comparisons; Post Hoc Comparisons; Scheffé Test; Tukey's Honestly Significant Difference (HSD)

### Further Readings

Abdi, H., Edelman, B., Valentin, D., & Dowling, W. J. (2009). *Experimental design and analysis for psychology.* Oxford, UK: Oxford University Press.

Hayter, A. J. (1986). The maximum familywise error rate of Fisher's least significant difference test. *Journal of the American Statistical Association, 81,* 1001–1004.

Seaman, M. A., Levin, J. R., & Serlin, R. C. (1991). New developments in pairwise multiple comparisons: Some powerful and practicable procedures. *Psychological Bulletin, 110,* 577–586.

## Fixed-Effects Models

Fixed-effects models are a class of statistical models in which the levels (i.e., values) of independent variables are assumed to be fixed (i.e., constant), and only the dependent variable changes in response to the levels of independent variables. This class of models is fundamental to the general linear models that underpin fixed-effects regression analysis and fixed-effects analysis of variance, or ANOVA (fixed-effects ANOVA can be unified with fixed-effects regression analysis by using dummy variables to represent the levels of independent variables in a regression model; see the article by Andrew Gelman for more information); the generalized linear models, such as logistic regression for binary response variables and binomial counts; Poison regression for Poisson (count) response variables; as well as the analysis of categorical data using such techniques as the Mantel-Haenszel or Peto odds ratio. A common thesis in assuming a fixed-effects model among these analyses is that under conditions of similar investigation methods, similar measurements, and similar experimental or observational units, the mean response among the levels of independent variables should be comparable. If there is any discrepancy, the difference is caused by the within-study variation among the effects at the fixed levels of independent variables. This entry discusses the application of fixed effects in designed experiments and observational studies, along with alternate applications.

### Designed Experiments

Fixed-effects models are very popular in designed experiments. The principal idea behind using these models is that the levels of independent variables (treatments) are specifically chosen by the researcher, whose sole interest is the response of the dependent

variable to the specific levels of independent variables that are employed in a study. If the study is to be repeated, the same levels of independent variables would be used again. As such, the inference space of the study, or studies, is the specific set of levels of independent variables. Results are valid only at the levels that are explicitly studied, and no extrapolation is to be made to levels of independent variables that are not explicitly investigated in the study.

In practice, researchers often arbitrarily and systematically choose some specific levels of independent variables to investigate their effects according to a hypothesis and/or some prior knowledge about the relationship between the dependent and independent variables. These levels of independent variables are either of interest to the researcher or thought to be representative of the independent variables. During the experiment, these levels are maintained constant. Measurements taken at each fixed level of an independent variable or a combination of independent variables therefore constitute a known population of responses to that level (combination of levels). Analyses then draw information from the mean variation of the study to make inference about the effect of those specific independent variables at the specified levels on the mean response of the dependent variable. A key advantage of a fixed-effects model design is that important levels of an independent variable can be purposefully investigated. As such, both human and financial resource utilization efficiency may be maximized. Examples of such purposeful investigations may be some specific dosages of a new medicine in a laboratory test for efficacy, or some specific chemical compositions in metallurgical research on the strength of alloy steel, or some particular wheat varieties in an agriculture study on yields.

The simplest example of a fixed-effects model design for comparing the difference in population means is the paired $t$ test model in a paired comparison design. This design is a variation of the more general randomized block design in that each experimental unit serves as a block. Two treatments are applied to each experimental unit, with the order varying randomly from one experiment unit to the next. The null hypothesis of the paired $t$ test is $\mu_1 - \mu_2 = 0$. Because of no sampling variability between treatments, the precision of estimates in this design is considerably improved as compared to a two-sample $t$ test model.

Fixed-effects model experiment designs contrast sharply with random effects model designs, in which the levels of independent variables are randomly selected from a large population of all possible levels. Either this population of levels is infinite in size, or the size is sufficiently large and can practically be considered infinite. The levels of independent variables are therefore believed to be random variables. Those that are chosen for a specific experiment are a random draw. If the experiment is repeated, these same levels are unlikely to be reused. Hence, it is meaningless to compare the means of the dependent variable at those specific levels in one particular experiment. Instead, the experiment seeks inference about the effect of the entire population of all possible levels, which are much broader than the specific ones used in the experiment, whether or not they are explicitly studied. In doing so, a random-effects model analysis draws conclusions from both within- and between-variable variation. Compared to fixed-effects model designs, the advantages of random-effects model designs are a more efficient use of statistical information, and results from one experiment can be extrapolated to levels that are not explicitly used in that experiment. A key disadvantage is that some important levels of independent variables may be left out of an experiment, which could potentially have an adverse effect on the generality of conclusions if those omitted levels turn out to be critical.

To illustrate the differences between a fixed- and a random-effects model analysis, consider a controlled, two-factor factorial design experiment. Assume that Factor A has $a$ levels; Factor B has $b$ levels; and $n$ measurements are taken from each combination of the levels of the two factors. Table 1 illustrates the ANOVA table comparing a fixed- and a random-effects model. Notice in Table 1 that the initial steps for calculating the mean squares are similar in both analyses. The differences are the expected mean squares and the construction of hypothesis tests. Suppose that the conditions for normality, linearity, and equal variance are all met, and the hypothesis tests on both the main and the interactive effects of the fixed-effects model ANOVA are simply concerned with the error variance, which is the expected mean square of the experimental error. In comparison, the hypothesis tests on the main effect of the random-effects model ANOVA draw information

**Table 1** Analysis of Variance Table for the Two-Factor Factorial Design Comparing a Fixed-With a Random-Effects Model , Where Factor A Has *a* Levels, Factor B Has *b* Levels, and *n* Replicates Are Measured at Each A × B Level

| *Source of Variance* | *Sum of Squares* | *Degrees of Freedom* | *Mean Square* | *Fixed-Effects Model* | | *Random-Effects Model* | |
|---|---|---|---|---|---|---|---|
| | | | | *Expected MS* | $F_0$ | *Expected MS* | $F_0$ |
| Factor A | $SS_A$ | $a-1$ | $MS_A = \frac{SS_A}{a-1}$ | $E(MS_B) = \sigma^2 + \frac{bn\sum_{i=1}^{a} A_i^2}{a-1}$ | $F_0 = \frac{MS_A}{MS_E}$ | $E(MS_A) = \sigma^2 + n\sigma^2_{AB} + bn\sigma^2_A$ | $F_0 = \frac{MS_A}{MS_{AB}}$ |
| Factor B | $SS_B$ | $b-1$ | $MS_B = \frac{SS_B}{b-1}$ | $E(MS_B) = \sigma^2 + \frac{an\sum_{j=1}^{b} B_j^2}{b-1}$ | $F_0 = \frac{MS_B}{MS_E}$ | $E(MS_B) = \sigma^2 + n\sigma^2_{AB} + an\sigma^2_B$ | $F_0 = \frac{MS_B}{MS_{AB}}$ |
| A × B | $SS_{AB}$ | $(a-1)(b-1)$ | $MS_{AB} = S\frac{S_{AB}}{(a-1)(b-1)}$ | $E(MS_{AB}) = \sigma^2 + \frac{n\sum_{i=1}^{a}\sum_{j=1}^{b}(AB)_{ij}^2}{(a-1)(b-1)}$ | $F_0 = \frac{MS_{AB}}{MS_E}$ | $E(MS_AB) = \sigma^2 + n\sigma^2_{AB}$ | $F_0 = \frac{MS_{AB}}{MS_E}$ |
| Error | $SS_E$ | $ab(n-1)$ | $MS_E = \frac{SS_E}{ab(n-1)}$ | $E(MS_E) = \sigma^2$ | | $E(MS_E) = \sigma^2$ | |
| Total | $SS_T$ | $abn-1$ | | | | | |

from both the experimental error variance and the variance due to the interactive effect of the main experimental factors. In other words, hypothesis tests in random-effects model ANOVA must be determined according to the expected mean squares. Finding an appropriate error term for a test is not as straightforward in a random-effects model analysis as in a fixed-effects model analysis, particularly when sophisticated designs such as the split-plot design are used. In these designs, one often needs to consult an authoritative statistical textbook, but not overly rely on commercial statistical software if he or she is not particularly familiar with the relevant analytical procedure.

## Observational Studies

In observational studies, researchers are often unable to manipulate the levels of an independent variable as they can frequently do in controlled experiments. Being the nature of an observational study, there may be many influential independent variables. Of them, some may be correlated with each other, whereas others are independent. Some may be observable, but others are not. Of the unobservable variables, researchers may have knowledge of some, but may be unaware of others. Unobservable variables are generally problematic and could complicate data analyses. Those that are hidden from the knowledge of the researchers are probably the worst offenders. They could potentially lead to erroneous conclusions by obscuring the main results of a study. If a study takes repeated measures (panel data), some of those variables may change values over the course of the study, whereas others may not. All of these add complexity to data analyses.

If an observational study does have panel data, the choice of statistical models depends on whether or not the variables in question are correlated with the main independent variables. The fixed-effects model is an effective tool if variables are correlated, whether they are measured or unmeasured. Otherwise, a random-effects model should be employed. Size of observation units (i.e., number of students in an education study) or groupings of such units (i.e., number of schools) is generally not a good criterion for choosing one particular statistical model over the other.

Take as an example a hypothetical ecological study in 10 cities of a country on the association between lung cancer prevalence rates and average cigarette consumption per capita in populations 45 years of age and older. Here, cigarette consumption is the primary independent variable and lung cancer prevalence rate is the dependent variable. Suppose that two surveys are done at two different times and noticeable differences are observed in both cigarette consumption and lung cancer prevalence rates at each survey time both across the 10 cities (i.e., intercity variation) and within each of the 10 cities (i.e., intracity variation over time). Both fixed and random regression models can be used to analyze the data, depending on the assumption that one makes.

A fixed-effects regression model can be used if one makes assumptions such as no significant changes in the demographic characteristics, in the cigarette supply-and-demand relationship, in the air pollution level and pollutant chemical composition, or other covariates that might be inductive to lung cancer over time in each city, and if one further assumes that any unobservable variable that might simultaneously affect the lung cancer prevalence rate and the average per capita cigarette consumption does not change over time.

$$y_{it} = \beta_0 + \beta_1 x_{it} + \alpha_i + \varepsilon_{it} \quad (1)$$

where $y_{it}$ and $x_{it}$ are, respectively, the lung cancer prevalence rate and the average per capita cigarette consumption in the $i$th city at time $t$; $\alpha_i$ is a fixed parameter for the $i$th city; and $\varepsilon_{it}$ is the error term for the $i$th city at time $t$. In this model, $\alpha_i$ captures the effects of all observed and unobserved time-invariant variables, such as demographic characteristics including age, gender, and ethnicity; socioeconomic characteristics; air pollution; and other variables, which could vary from city to city but are constant within the $i$th city (this is why the above model is called a fixed-effects model). By treating $\alpha_i$ as fixed, the model focuses only on the within-city variation while ignoring the between-city variation.

The estimation of Equation 1 becomes inefficient if many dummy variables are included in the model to accommodate a large number of observational units ($\alpha_i$) in panel data, because this sacrifices many degrees of freedom. Furthermore,

a large number of observational units coupled with only a few time points may result in the intercepts of the model containing substantial random error, making them inconsistent. Not much, if any, information could be gained from those noisy parameters. To circumvent these problems, one may convert the values of both the dependent and the independent variables of each observational unit into the difference from their respective mean for that unit. The differences in the dependent variable are then regressed on the differences of the independent variables without an intercept term. The estimator then looks only at how changes in the independent variables cause the dependent variable to vary around a mean within an observational unit. As such, the unit effects are removed from the model by differencing.

It is clear from the above discussions that the key technique in using a fixed-effects model in panel data is to allow each observational unit ("city" in the earlier example) to serve as its own control so that the data are grouped. Consequently, a great strength of the fixed-effects model is that it simultaneously controls for both observable and unobservable variables that are associated with each specific observational unit. The fixed-effect coefficients ($\alpha_i I$) absorb all of the across-unit influences, leaving only the within-unit effect for the analysis. The result then simply shows how much the dependent variable changes, on average, in response to the variation in the independent variables within the observational units; that is, in the earlier example, how much, on average, the lung cancer prevalence rate will go up or down in response to each unit change in the average cigarette consumption per capita.

Because fixed-effects regression model analyses depend on each observational unit serving as its own control, key requirements in applying them in research are as follows: (a) There must be two or more measurements on the same dependent variable in an observational unit; otherwise, the unit effect cannot be properly controlled; and (b) independent variables of interest must change values on at least two of the measurement occasions in some of the observational units. In other words, the effect of any independent variable that does not have much within-unit variation cannot be estimated. Observational units with values of little within-unit variation in some independent variables contribute less information to the overall analysis with respect to those variables of concern. The second point is easy to understand if one treats a variable that does not change values as a constant. A constant subtracting a constant is zero—that is, a zero effect of such variables on the dependent variable. In this regard, fixed-effects models are mostly useful for studying the effects of independent variables that show within-observational-unit variation.

If, on the other hand, there is reasonable doubt regarding the assumptions made about a fixed-effects model, particularly if some independent variables are not correlated with the major independent variable(s), a fixed-effects model will not be able to remove the bias caused by those variables. For instance, in the above hypothetical study, a shortage in the cigarette supply caused a decrease in its consumption in some cities, or a successful promotion by cigarette makers or retailers persuaded more people to smoke in other cities. These random changes from city to city make a fixed-effects model unable to control effectively the between-city variation in some of the independent variables. If this happens, a random-effects model analysis would be more appropriate because it is able to accommodate the variation by incorporating in the model two sources of error. One source is specific to each individual observational unit, and the other source captures variation both within and between individual observational units.

## Alternate Applications

After discussing the application of fixed-effects model analyses in designed and observational research, it may also be helpful to mention the utility of fixed-effects models in meta-analysis (a study on studies). This is a popular technique widely used for summarizing knowledge from individual studies in social sciences, health research, and other scientific areas that rely mostly on observational studies to gather evidence. Meta-analysis is needed because both the magnitude and the direction of the effect size could vary considerably among observational studies that address the same question. Public policies, health practices, or products developed based on the result of each individual study therefore may not be able to achieve their desired effects as designed or as believed. Through meta-analysis, individual studies are

brought together and appraised systematically. Common knowledge is then explicitly generated to guide public policies, health practices, or product developments.

In meta-analysis, each individual study is treated as a single analysis unit and plugged into a suitable statistical model according to some assumptions. Fixed-effects models have long been used in meta-analysis, with the following assumptions: (a) Individual studies are merely a sample of the same population, and the true effect for each of them is therefore the same; and (b) there is no heterogeneity among study results. Under these assumptions, only the sampling error (i.e., the within-study variation) is responsible for the differences (as reflected in the confidence interval) in the observed effect among studies. The between-study variation in the estimated effects has no consequence on the confidence interval in a fixed-effects model analysis. These assumptions may not be realistic in many instances and are frequently hotly debated. An important difficulty in applying a fixed-effects model in meta-analysis is that each individual study is conducted on different study units (individual persons, for instance) under a different set of conditions by different researchers. Any (or all) of these differences could potentially introduce its (or their) effects into the studies to cause variation in their results. Therefore, one needs to consider not only within- but also between-study variation in a model in order to generalize knowledge properly across studies. Because the objective of meta-analyses is to seek validity generalization, and because heterogeneity tests are not always sufficiently sensitive, a random-effects model is thus believed to be more appropriate than a fixed-effects model. Unless there is truly no heterogeneity confirmed through proper investigations, fixed-effects model analyses tend to overestimate the true effect by producing a smaller confidence interval. On the other hand, critics argue that random-effects models make assumptions about distributions, which may or may not be realistic or justified. They give more weight to small studies and are more sensitive to publication bias. Readers interested in meta-analysis should consult relevant literature before embarking on a meta-analysis mission.

The arguments for and against fixed- and random-effects models seem so strong, at least on the surface, that a practitioner may be bewildered in the task of choosing the right model for specific research. In ANOVA, after a model is chosen, there is no easy way to identify the correct variance components for computation of standard errors and for hypothesis tests (see Table 1, for example). This leads Gelman to advocate abolishing the terminology of fixed- and random-effects models. Instead, a unified approach is taken within a hierarchical (multilevel) model framework, regardless of whether one is interested in the effects of specific treatments used in a particular experiment (fixed-effects model analyses in a traditional sense) or in the effects of the underlying population of treatments (random-effects model analyses otherwise). In meta-analysis, Bayesian model averaging is another alternative to fixed- and random-effects model analyses.

## Final Thoughts

Fixed-effects models concern mostly the response of dependent variables at the fixed levels of independent variables in a designed experiment. Results thus obtained generally are not extrapolated to other levels that are not explicitly investigated in the experiment. In observational studies with repeated measures, fixed-effects models are used principally for controlling the effects of unmeasured variables if these variables are correlated with the independent variables of primary interest. If this assumption does not hold, a fixed-effects model cannot adequately control for inter-unit variation in some of the independent variables. A random-effects model would be more appropriate.

*Shihe Fan*

***See also*** Analysis of Variance (ANOVA); Bivariate Regression; Random-Effects Models

### Further Readings

Gelman, A. (2005). Analysis of variance: Why it is more important than ever. *Annals of Statistics, 33,* 1–53.

Hocking, R. R. (2003). *Methods and application of linear models: Regression and the analysis of variance* (2nd ed.). Hoboken, NJ: Wiley.

Kuehl, R. O. (1994). *Statistical principles of research design and analysis.* Belmont, CA: Duxbury.

Montgomery, D. C. (2001). *Design and analysis of experiments* (5th ed.). Toronto: Wiley.

Nelder, J. A., & Weddenbum, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A, 135,* 370–384.
Sutton, A. J., Abrams, K. R., Jones, D. R., Sheldon, T. A., & Song, F. (2000). *Methods for meta-analysis in medical research*. Toronto, Ontario, Canada: Wiley.

# Focus Group

A focus group is a form of qualitative research conducted in a group interview format. The focus group typically consists of a group of participants and a researcher who serves as the moderator for discussions among the group members. In focus groups, there is not always the usual exchange of questions and answers between the researcher and the group that one would commonly envision in an interview setting. Rather, the researcher often ensures that specific topics of research interest are discussed by the entire group in hopes of extracting data and self-disclosure that might otherwise be withheld in the traditional researcher-interviewee environment. In this entry, the purpose, history, format, advantages and disadvantages, and future direction of focus groups are discussed.

## Purpose

The purpose of focus groups varies depending on the topic under investigation. In some studies, a focus group serves as the primary means of collecting data via a strictly qualitative approach. In other studies, a group discussion or focus group is used as a preliminary step before proceeding to quantitative data collection, usually known as a mixed-method approach. In still other studies, a focus group is employed in conjunction with individual interviews, participant observation, and other qualitative forms of data collection. Thus, focus groups are used to complement a mixed-method study or they have a self-contained function, given their ability to function independently or be combined with other qualitative or quantitative approaches. As a result of their versatility, focus groups serve the needs of many researchers in the social sciences.

## History

The history of focus groups dates back to the 1930s, when Emory S. Bogardus, a scholar, wrote about group interviews and their usefulness to researchers. During World War II, focus groups were conducted to determine the usefulness of the military's training materials and the success of war propaganda in the war effort. Following World War II, focus groups were used primarily to obtain responses and gather opinions about films, written materials, and radio broadcasts. Beginning in the 1980s, focus groups were used in a wide variety of research settings, thus expanding their initial role as mere gauges for government and marketing research.

## Format

Although focus groups have enormous versatility and diversity in how they operate, a step-by-step format for conducting focus groups has emerged in recent years. The first step is to determine the goals of the study. Although not highly specific at this point, it is common for the researcher to write a general purpose statement that lays the foundation for the research project. The second step is to determine who will serve as the moderator of the focus group. Selection of a moderator is of utmost importance to the success of the focus group, as the moderator promotes interactions among group members and prevents the group from digressing from the topic of interest. The next step involves refinement of research goals. Lists of information to obtain during the focus group interviews are created, and these lists serve as the basis for formulating questions and probes to be used later in the focus group interviews. Following this step, participants are recruited for the focus group, preferably through intentional sampling, with the goal of obtaining a group of individuals that is most apt to provide the researcher with the needed information. After the participants have been selected, the number of focus group sessions is determined. The number of focus group sessions will vary and will depend on the number of participants needed to make the focus group interviews successful. The next step is to locate a focus group

site where the interviews will be conducted. However, there are no specific parameters for how this step is to be accomplished. The seventh step involves the development of an interview guide. The interview guide includes the research objectives and ensuing questions that have been developed and refined from earlier steps in the process. The questions are constructed with the intent to facilitate smooth transitions from one topic to another. The culminating step is to conduct the focus group interview. The moderator should be well prepared for the focus group session. Preparation includes having the necessary documents, arranging the room and chairs, and arriving early to the site to test any media to be used during the focus group interview. After the group interview is conducted and recorded, it is most frequently transcribed, coded, and analyzed. The interview is transcribed so the researcher is able to adequately interpret the information obtained from the interview.

### Advantages and Disadvantages

Focus groups are created to address the researcher's needs, and there are certain advantages and disadvantages inherent in their use. One advantage associated with the use of focus groups is that they are time-efficient in comparison to the traditional one-on-one interviews. The ability to collect information from multiple individuals at one time instead of interviewing one person at a time is an attractive option to many researchers. However, there are some disadvantages associated with the group format. Participants might find the time to travel to the focus group facility site to be burdensome, and the moderator might find scheduling a time for the focus group to meet to be extremely challenging.

Another advantage of the use of focus groups lies in their dependence on group interactions. With multiple individuals in one setting, discussions and the infusion of diverse perspectives during those discussions are possible. However, discussions of any focus group are dependent upon group dynamics, and the data gleaned from those discussions might not be as useful if group members are not forthcoming or an uncomfortable mood becomes evident and dominates a session.

### Future Directions

Focus groups are held in a variety of settings, such as a formal setting where psychologists do research or a less structured environment where a book club meets to discuss and react to a novel. For decades, focus groups have served governments, researchers, businesses, religious groups, and many other areas of society. Thus, focus groups are likely to continue to be used in the future when individuals and researchers are in need of qualitative data.

*Matthew J. Grumbein and*
*Patricia A. Lowe*

***See also*** Qualitative Research

#### Further Readings

Kormanski, C. (1999). *The team: Explorations in group process.* Denver, CO: Love.
Moore, C. M. (1987). *Group techniques for idea building.* Newbury Park, CA: Sage.
Morgan, D. L. (1997). *Qualitative research methods* (2nd ed.). Thousand Oaks, CA: Sage.
Vaughn, S., Schumm, J. S., & Sinagub, J. (1996). *Focus group interviews in education psychology.* Thousand Oaks, CA: Sage.

## Follow-Up

Follow-up procedures are an important component of all research. They are most often conducted during the actual research but can also be conducted afterward. Follow-up is generally done to increase the overall effectiveness of the research effort. It can be conducted for a number of reasons, namely, to further an end in a particular study, review new developments, fulfill a research promise, comply with institutional review board protocol for research exceeding a year, ensure that targeted project milestones are being met, thank participants or informants for their time, debrief stakeholders, and so on. Follow-up may also be conducted as a normal component of the research design. Or, it could even be conducted subsequent to the original research to ascertain if an intervention has

changed the lives of the study participants. Regardless of its purpose, follow-up always has cost implications.

## Typical Follow-Up Activities

### *Participants*

In the conduct of survey research, interviewers often have to make multiple attempts to schedule face-to-face and telephone interviews. When face-to-face interviews are being administered, appointments generally need to be scheduled in advance. However, participants' schedules may make this simple task difficult. In some cases, multiple telephone calls and/or letters may be required in order to set up a single interview. In other cases (e.g., national census), follow-up may be required because participants were either not at home or were busy at the time of the interviewer's visits. Likewise, in the case of telephone interviews, interviewers may need to call potential participants several times before they are actually successful in getting participants on the phone.

With mail surveys, properly timed (i.e., predefined follow-up dates—usually every 2 weeks) follow-up reminders are an effective strategy to improve overall response rates. Without such reminders, mail response rates are likely to be less than 50%. Follow-up reminders generally take one of two forms: a letter or postcard reminding potential participants about the survey and encouraging them to participate, or a new survey package (i.e., a copy of the survey, return envelope, and a reminder letter). The latter technique generally proves to be more effective because many potential participants either discard mail surveys as soon as they are received or are likely to misplace the survey if it is not completed soon after receipt.

### *Review New Developments*

During a particular research study, any number of new developments can occur that would require follow-up action to correct. For example, a pilot study may reveal that certain questions were worded in such an ambiguous manner that most participants skipped the questions. To correct this problem, the questions would need to be reworded and a follow-up pilot study would need to be administered to ascertain clarity of the reworded questions. Likewise, a supervisor may discover that one or more telephone interviewers are not administering their telephone surveys according to protocol. This would require that some follow-up training be conducted for those interviewers.

### *Project Milestones*

Research activities require careful monitoring and follow-up to ensure that things are progressing smoothly. Major deviations from project milestones generally require quick follow-up action to get the activity back on schedule to avoid schedule slippage and cost overruns.

### *Incentives*

In research, incentives are often offered to encourage participation. Researchers and research organizations therefore need to follow up on their promises and mail the promised incentive to all persons who participated in the research.

### *Thank-You Letters*

Information for research is collected using a number of techniques (e.g., focus groups, informants, face-to-face interviews). Follow-up thank-you letters should be a normal part of good research protocol to thank individuals for their time and contributions.

### *Stakeholder Debriefing*

Following the completion of the research, one or more follow-up meetings may be held with stakeholders to discuss the research findings, as well as any follow-up studies that may be required.

### *Compliance With Institutional Review Boards*

The U.S. Department of Health and Human Services (Office of Human Research Protections) Regulation 45 CFR 46.109(e) requires that institutional review boards conduct follow-up reviews at least annually on a number of specific issues when research studies exceed one year.

### *Follow-Up Studies*

Follow-up studies may be a component of a particular research design. For example, time series designs include a number of pretests and posttests using the same group of participants at different intervals. If the purpose of the posttest is to ascertain the strength of a particular treatment over an extended period, the posttest is referred to as follow-up. Follow-up studies may also be conducted when cost and time are constraining factors that make longitudinal studies unfeasible. For example, a follow-up study on the same participants can be held at the end of a 20-year period, rather than at the end of every 5-year period.

### *Success of Intervention*

In some types of research, follow-up may be conducted subsequent to the original research to ascertain if an intervention has changed the lives of the study participants and to ascertain the impact of the change.

### Cost Implications

All follow-up activities have associated costs that should be estimated and included in project budgets. The extent and type of follow-up will, to a large extent, determine the exact cost implications. For example, from a cost-benefit standpoint, mail survey follow-up should be limited to three or four repeat mailings. In addition, different costs will be incurred depending on the follow-up procedure used. For example, a follow-up letter or postcard will cost a lot less (less paper, less weight, less postage) than if the entire survey package is reposted. When mail surveys are totally anonymous, this will also increase costs because the repeat mailings will have to be sent to all participants. Using certified mail generally improves response rates, but again at additional cost. When thank-you letters are being sent, they should be printed on official letterhead and carry an official signature if possible. However, the use of official letterheads and signatures is more costly compared to a computer printout with an automated signature. In the final analysis, researchers need to carefully balance costs versus benefits when making follow-up decisions.

*Nadini Persaud*

*See also* Debriefing; Interviewing; Recruitment; Survey

**Further Readings**

Babbie, E. (2004). *The practice of social research* (10th ed.). Belmont, CA: Wadsworth/Thompson Learning.

Office of Human Research Protection. (2007, January). Guidance on continuing review [Online]. Retrieved April 19, 2009, from http://www.hhs.gov/ohrp/humansubjects/guidance/contrev0107.htm

## Frequency Distribution

A frequency distribution shows all the possible scores a variable has taken in a particular set of data, together with the frequency of occurrence of each score in the respective set. This means that a frequency distribution describes how many times a score occurs in the data set.

Frequency distributions are one of the most common methods of displaying the pattern of observations for a given variable. They offer the possibility of viewing each score and its corresponding frequency in an organized manner within the full range of observed scores. Along with providing a sense of the most likely observed score, they also show, for each score, how common or uncommon it is within the analyzed data set.

Both discrete and continuous variables can be described using frequency distributions. Frequency distributions of a particular variable may be displayed using stem-and-leaf plots, frequency tables, and frequency graphs (typically bar charts or histograms, and polygons). This entry discusses each of these types of displays, along with its shape and modality, and the advantages and drawbacks of using frequency distributions.

### Stem-and-Leaf Plots

Stem-and-leaf plots were developed by John Tukey in the 1970s. To create a stem-and-leaf plot for a set of data, the raw data first must be arranged in an array (in ascending or descending order). Then, each number must be separated into a stem and a leaf. The stem consists of the first digit or

digits, and the leaf consists of the last digit. Whereas the stem can have any number of digits, the leaf will always have only one. Table 1 shows a stem-and-leaf plot of the ages of the participants at a city hall meeting.

The plot shows that 20 people have participated at the city hall meeting, five in their 30s, none in his or her 40s, eight in their 50s, five in their 60s, and two in their 70s.

Stem-and-leaf plots have the advantage of being easily constructed from the raw data. Whereas the construction of cumulative frequency distributions and histograms often requires the use of computers, stem-and-leaf plots are a simple paper-and-pencil method for analyzing data sets. Moreover, no information is lost in the process of building up stem-and-leaf plots, as is the case in, for example, grouped frequency distributions.

## Frequency Tables

A table that shows the distribution of the frequency of occurrence of the scores a variable may take in a data set is called a frequency table. Frequency tables are generally univariate, because it is more difficult to build up multivariate tables. They can be drawn for both ungrouped and grouped scores. Frequency tables with ungrouped scores are typically used for discrete variables and when the number of different scores the variable may take is relatively low. When the variable to be analyzed is continuous and/or the number of scores it may take is high, the scores are usually grouped into classes.

Two steps must be followed to build a frequency table out of a set of data. First, the scores or classes are arranged in an array (in an ascending or descending order). Then, the number of observations corresponding to each score or falling within each class is counted. Table 2 presents a frequency distribution table for the age of the participants at the city hall meeting from the earlier example.

Apart from a list of the scores or classes and their corresponding frequencies, frequency tables may also contain relative frequencies or proportions (obtained by dividing the simple frequencies by the number of cases) and percentage frequencies (obtained by multiplying the relative frequencies by 100).

Frequency tables may also include cumulative frequencies, proportions, or percentages. Cumulative frequencies are obtained by adding the frequency of each observation to the sum of the frequencies of all previous observations. Cumulative proportions and cumulative percentages are calculated similarly; the only difference is that, instead of simple frequencies, cumulative frequencies are divided by the total number of cases for obtaining cumulative proportions.

Frequency tables look similar for nominal or categorical variables, except the first column contains categories instead of scores or classes. In some frequency tables, the missing scores for nominal variables are not counted, and thus, proportions and percentages are computed based on the number of nonmissing scores. In other frequency tables, the missing scores may be included as a category so that proportions and percentages can be computed based on the full sample size of nonmissing and missing scores. Either approach has analytical value, but authors must be clear about which base number is used in calculating any proportions or percentages.

**Table 1** Stem-and-Leaf Plot of the Ages of the Participants at a City Hall Meeting

| *Stem* | *Leaf* |
|---|---|
| 3 | 34457 |
| 4 | |
| 5 | 23568889 |
| 6 | 23778 |
| 7 | 14 |

**Table 2** Frequency Table of the Age of the Participants at a City Hall Meeting

| *Age y* | *Frequency f* | *Relative Frequency* $rf = f/n$ | *Percentage Frequency* $p = 100 * rf$ |
|---|---|---|---|
| 30–39 | 5 | 0.25 | 25.00 |
| 40–49 | 0 | 0.00 | 00.00 |
| 50–59 | 8 | 0.40 | 40.00 |
| 60–69 | 5 | 0.25 | 25.00 |
| 70–79 | 2 | 0.10 | 10.00 |
| | $n = 20$ | 1.00 | 100.00% |

## Frequency Graphs

Frequency graphs can take the form of bar charts and histograms, or polygons.

### *Bar Charts and Histograms*

A frequency distribution is often displayed graphically through the use of a bar chart or a histogram. Bar charts are used for categorical variables, whereas histograms are used for scalable variables. Bar charts resemble histograms in that bar heights correspond to frequencies, proportions, or percentages. Unlike the bars in a histogram, bars in bar charts are separated by spaces, thus indicating that the categories are in arbitrary order and that the variable is categorical. In contrast, spaces in a histogram signify zero scores.

Both bar charts and histograms are represented in an upper-right quadrant delimited by a horizontal $x$-axis and a vertical $y$-axis. The vertical axis typically begins with zero at the intersection of the two axes; the horizontal scale need not begin with zero, if this leads to a better graphic representation. Scores are represented on the horizontal axis, and frequencies, proportions, or percentages are represented on the vertical axis. When working with classes, either limits or midpoints of class interval are measured on the $x$-axis. Each bar is centered on the midpoint of its corresponding class interval; its vertical sides are drawn at the real limits of the respective interval. The base of each bar represents the width of the class interval.

A frequency distribution is graphically displayed on the basis of the frequency table that summarizes the sample data. The frequency distribution in Table 2 is graphically displayed in the histogram depicted in Figure 1.

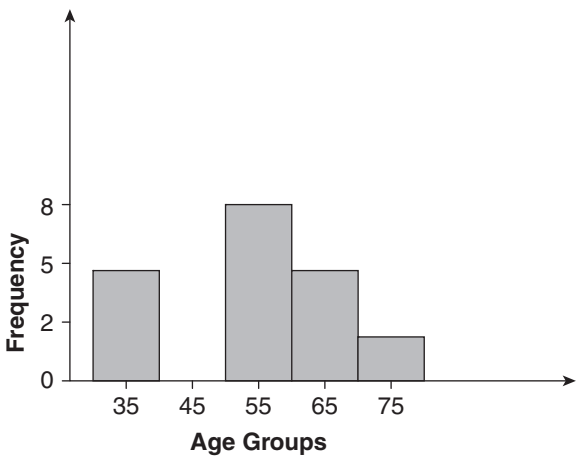



**Figure 1** Histogram of the Age of the Participants at a City Hall Meeting

### *Frequency Polygons*

Frequency polygons are drawn by joining the points formed by the midpoint of each class interval and the frequency corresponding to that class interval. However, it may be easier to derive the frequency polygon from the histogram. In this case, the frequency polygon is drawn by joining the midpoints of the upper bases of adjacent bars of the histogram by straight lines. Frequency polygons are typically closed at each end. To close them, lines are drawn from the point given by the midpoint of the upper base of each of the histogram's end columns to the midpoints of the adjacent intervals (on the $x$-axis). Figure 2 presents a frequency polygon based on the histogram in Figure 1.

Histograms and frequency polygons may also be constructed for relative frequencies and percentages in a similar way. The advantage of using graphs of relative frequencies is that they can be used to directly compare samples of different sizes. Frequency polygons are especially useful for graphically depicting cumulative distributions.

## Distribution Shape and Modality

Frequency distributions can be described by their skewness, kurtosis, and modality.

### *Skewness*

Frequency distributions may be symmetrical or skewed. Symmetrical distributions imply equal proportions of cases at any given distance above and below the midpoint on the score range scale. Consequently, each half of a symmetrical distribution looks like a mirror image of the other half. Symmetrical distributions may be uniform (rectangular) or bell-shaped, and they may have one, two, or more peaks.

Perfectly symmetrical distributions are seldom encountered in practice. Skewed or asymmetrical

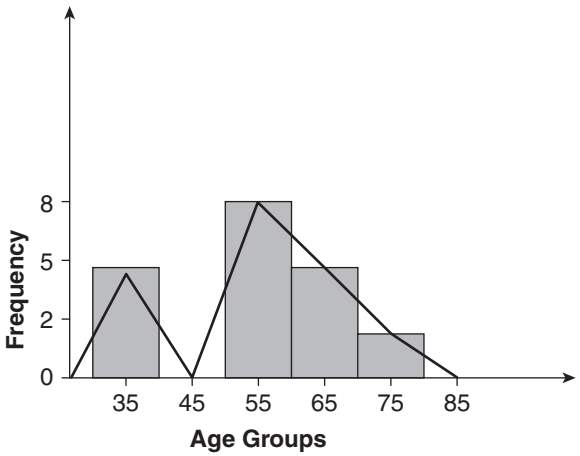



**Figure 2** Frequency Polygon of the Age of the Participants at a City Hall Meeting

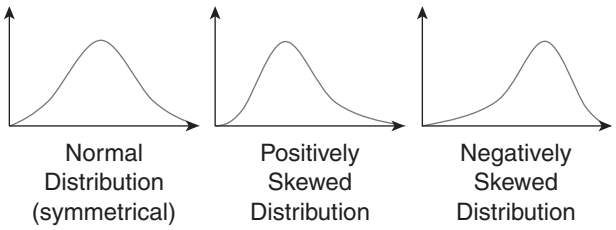



**Figure 3** Example of Symmetrical, Positively Skewed, and Negatively Skewed Distributions

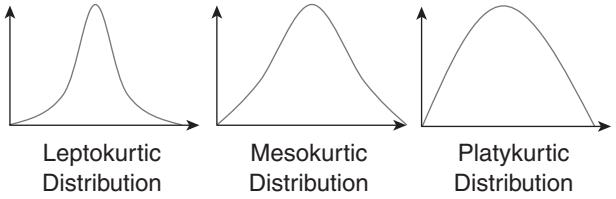



**Figure 4** Example of Leptokurtic, Mesokurtic, and Platykurtic Distributions

distributions, which are not symmetrical and typically have protracted tails, are more common. This means that scores are clustered at one tail of the distribution, while occurring less frequently at the other tail. Distributions are said to be positively skewed if scores form a tail to the right of the mean, or negatively skewed if scores form a tail to the left of the mean. Positive skews occur frequently in variables that have a lower limit of zero, but no specific upper limit, such as income, population density, and so on.

The degree of skewness may be determined by using a series of formulas that produces the index of skewness; a small index number indicates a small degree of skewness, and a large index number indicates significant nonsymmetry. Truly symmetrical distributions have zero skewness. Positive index numbers indicate tails toward higher scores; negative index numbers indicate the opposite, tails toward lower scores. Generally, severely skewed distributions show the presence of outliers in the data set. Statistics computed using severely skewed data may be unreliable. Figure 3 presents symmetrical, positively skewed, and negatively skewed distributions.

### *Kurtosis*

Kurtosis refers to a frequency distribution's degree of flatness. A distribution may have the same degree of skewness, but differ in terms of the kurtosis. According to their degree of kurtosis, distributions may be leptokurtic, mesokurtic, or platykurtic. Leptokurtic distributions are characterized by a high degree of peakedness. These distributions are usually based on data sets in which most of the scores are grouped around the mode. Normal (Gaussian) distributions are described as mesokurtic. Platykurtic distributions are characterized by a higher degree of flatness, which means that more scores in the data set are distributed further away from the mode, as compared to leptokurtic and mesokurtic distributions. Figure 4 presents leptokurtic, mesokurtic, and platykurtic distributions.

### *Modality*

The mode is the most frequently occurring score in a distribution. Distributions may be unimodal (they have only one peak), bimodal (they have two peaks), or multimodal (they have more than two peaks). A distribution may be considered bimodal or multimodal even if the peaks do not represent scores with equal frequencies. Rectangular distributions have no mode. Multiple peaks may also occur in skewed distributions; in such cases, they may indicate that two or more dissimilar kinds of

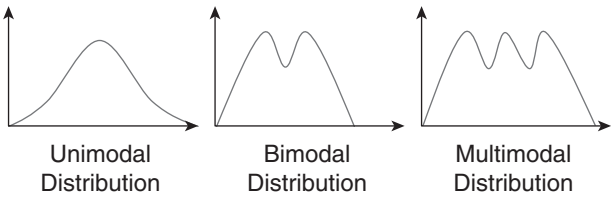



**Figure 5** Example of Unimodal, Bimodal, and Multimodal Distributions

cases have been combined in the analyzed data set. Multipeaked distributions are more difficult to interpret, resulting in misleading statistics. Distributions with multiple peaks suggest that further research is required in order to identify subpopulations and determine their individual characteristics. Figure 5 presents unimodal, bimodal, and multimodal distributions.

### Advantages and Drawbacks of Using Frequency Distributions

The advantage of using frequency distributions is that they present raw data in an organized, easy-to-read format. The most frequently occurring scores are easily identified, as are score ranges, lower and upper limits, cases that are not common, outliers, and total number of observations between any given scores.

The primary drawback of frequency distributions is the loss of detail, especially when continuous data are grouped into classes and the information for individual cases is no longer available. The reader of Table 2 learns that there are eight participants 50 to 59 years old, but the reader does not receive any more details about individual ages and how they are distributed within this interval.

*Oana Pusa Mihaescu*

***See also*** Cumulative Frequency Distribution; Descriptive Statistics; Distribution; Frequency Table; Histogram

#### Further Readings

Downie, N. M., & Heath, R. W. (1974). *Basic statistical methods*. New York: Harper & Row.

Fielding, J. L., & Gilbert, G. N. (2000). *Understanding social statistics*. Thousand Oaks, CA: Sage.

Fried, R. (1969). *Introduction to statistics*. New York: Oxford University Press.

Hamilton, L. (1996). *Data analysis for social sciences: A first course in applied statistics*. Belmont, CA: Wadsworth.

Kiess, H. O. (2002). *Statistical concepts for the behavioral sciences*. Boston: Allyn & Bacon.

Kolstoe, R. H. (1973). *Introduction to statistics for the behavioral sciences*. Homewood, IL: Dorsey.

Lindquist, E. F. (1942). *A first course in statistics: Their use and interpretation in education and psychology*. Cambridge, MA: Riverside Press.

## Frequency Table

Frequency is a measure of the number of occurrences of a particular score in a given set of data. A frequency table is a method of organizing raw data in a compact form by displaying a series of scores in ascending or descending order, together with their frequencies—the number of times each score occurs in the respective data set. Included in a frequency table are typically a column for the scores and a column showing the frequency of each score in the data set. However, more detailed tables may also contain relative frequencies (proportions) and percentages. Frequency tables may be computed for both discrete and continuous variables and may take either an ungrouped or a grouped format. In this entry, frequency tables for ungrouped and grouped formats are discussed first, followed by a discussion of limits and midpoints. This entry concludes with a brief discussion of the advantages and drawbacks of using frequency tables.

### Frequency Tables for Distributions With Ungrouped Scores

Frequency distributions with ungrouped scores are presented in tables showing the scores in the first column and how often each score has occurred (the frequency) in the second. They are typically used for discrete variables, which have a countable or finite number of distinct values. Tables of ungrouped scores are also used when the number of different scores a variable can take in a data set is low.

**Table 1** Raw Data of the Number of Children Families Have in a Small Community

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 5 | 2 | 3 | 4 | 5 | 5 | 4 | 5 | 2 | 0 |

**Table 2** An Ascending Array of the Number of Children Families Have in a Small Community

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 2 | 2 | 3 | 4 | 4 | 5 | 5 | 5 | 5 |

**Table 3** Frequency Table of the Number of Children Families Have in a Small Community

| *Number of Children y* | *Frequency f* | *Relative Frequency* $rf = f/n$ | *Percentage Frequency* $p = 100 * rf$ |
|---|---|---|---|
| 0 | 1 | 0.10 | 10.00 |
| 1 | 0 | 0.00 | 00.00 |
| 2 | 2 | 0.20 | 20.00 |
| 3 | 1 | 0.10 | 10.00 |
| 4 | 2 | 0.20 | 20.00 |
| 5 | 4 | 0.40 | 40.00 |
| | $n = 10$ | 1.00 | 100.00% |

Two steps must be followed to build up a frequency table out of a set of data: (a) Construct a sensible array using the given set of data, and (b) count the number of times each score occurs in the given data set. The raw data in Table 1 show the number of children families have in a small community.

Building up an array implies arranging the scores in an ascending or descending order. An ascending array is built for the data set in Table 2.

The number of times each score occurs in the data set is then counted, and the total is displayed for each score, as in Table 3.

Frequencies measure the number of times each score occurs. This means that one family has no children, and four families have five children each. Although some scores may not occur in the sample data, these scores must nevertheless be listed in the table. For example, even if there are no families with only one child, the score of 1 is still displayed together with its corresponding frequency (zero) in the ascending array built out of the sample data.

Relative frequencies, also called proportions, are computed as frequencies divided by the sample size: $rf = f/n$. In this equation, $rf$ represents the relative frequency corresponding to a particular score, $f$ represents the frequency corresponding to the same score, and $n$ represents the total number of cases in the analyzed sample. They indicate the proportion of observations corresponding to each score. For example, the proportion of families with two children in the analyzed community is 0.20.

Percentages are computed as proportions multiplied by 100: $p = rf(100)$, where $p$ represents the percentage and $rf$ represents the relative frequency corresponding to a particular score. They indicate what percentage of observations corresponds to each score. For example, 20% of the families in the observed sample have four children each. Proportions in a frequency table must sum to 1.00, whereas percentages must sum to 100.00. Due to rounding off, some imprecision may sometimes occur, and the total proportion and percentage may be just short of or a little more than 1.00 or 100.00%, respectively. However, this issue is now completely solved through the use of computer programs for such calculations.

## Frequency Tables for Grouped Frequency Distributions

In grouped frequency distributions, the scores are organized into classes, typically arranged in ascending or descending order. The frequency of the observations falling into each class is recorded. Tables with grouped frequency distributions are typically used for continuous variables, which can take on an infinite number of scores. Grouped frequencies may be employed for discrete variables as well, especially when they take on too many scores to be represented in an ungrouped form. In this case, representing the frequency of each score does not offer many useful insights. Because of the wide range of scores in these data, it may be difficult to perceive any interesting characteristics of the analyzed data set.

To build up a grouped frequency table, a number of classes or groups are formed based on the sample data. A class is a range of scores into which raw scores are grouped. In this case, the frequency is not the number of times each score occurs, but the number of times these scores fall into one of these

**Table 4** Raw Data on Students' Test Scores

| | | |
|---|---|---|
| 86.5 | 87.7 | 78.8 |
| 88.1 | 86.2 | 87.3 |
| 99.6 | 96.3 | 92.1 |
| 92.5 | 79.1 | 89.8 |
| 76.1 | 98.8 | 98.1 |

classes. Four steps must be followed to build a grouped frequency table: (a) Arrange the scores into an array, (b) determine the number of classes, (c) determine the size or width of the classes, and (d) determine the number of observations that fall into each class.

Table 4 displays a sample of the scores obtained by a sample of students at an exam.

A series of rules applies to class selection: (a) All observations must be included; (b) each observation must be assigned to only one class; (c) no scores can fall between two intervals; and (d) whenever possible, class intervals (the width of each class) must be equal. Typically, researchers choose a manageable number of classes. Too few classes lead to a loss of information from the data, whereas too many classes lead to difficulty in analyzing and understanding the data. It is sometimes recommended that the width of the class intervals be determined by dividing the difference between the largest and the smallest score (the range of scores) by the number of class intervals to be used. Referring to the above example, students' lowest and highest test scores are 76.1 and 99.6, respectively. The width of the class interval, $i$, would then be found by computing: $i = \frac{99.6 - 76.1}{5} = 4.7$, if the desired number of classes is five. The five classes would then be: 76.1–80.8; 80.9–85.5; 85.6–90.2; 90.3–94.9; 95.0–99.6. However, this is not a very convenient grouping. It would be easier to use intervals of 5 or 10, and limits that are multiples of 5 or 10. There are many situations in which midpoints are used for analysis, and midpoints of 5 and 10 intervals are easier to calculate than midpoints of 4.7 intervals. Based on this reasoning, the observations in the example data set are grouped into five classes, as in Table 5. Finally, the number of observations that fall within each interval is counted.

Frequencies measure the number of cases that fall within each class. This means that three students have scored between 75.1 and 80.0, and four students have scored between 95.1 and the maximum score of 100.0. Again, there may be some classes for which the frequency is zero, meaning that no case falls within that class. However, these classes must also be listed (for example, no students have obtained between 80.1 and 85.0 points; nevertheless, this case is listed in Table 5).

**Table 5** Grouped Frequency Table of Students' Test Scores

| *Students' Test Scores y* | *Frequency f* | *Relative Frequency* $rf = f/n$ | *Percentage Frequency* $p = 100 * rf$ |
|---|---|---|---|
| 75.1–80.0 | 3 | 0.20 | 20.00 |
| 80.1–85.0 | 0 | 0.00 | 00.00 |
| 85.1–90.0 | 6 | 0.40 | 40.00 |
| 90.1–95.0 | 2 | 0.13 | 13.33 |
| 95.1–100.0 | 4 | 0.27 | 26.67 |
| | $n = 15$ | 1.00 | 100.00% |

In general, grouped frequency tables include a column displaying the classes and a column showing their corresponding frequencies, but they may also include relative frequencies (proportions) and percentages. Proportions and percentages are computed in the same way as for ungrouped frequency tables. Their meaning changes, though. For example, Table 5 shows that the proportion of students who have scored between 85.1 and 90.0 is 0.40, and that 13.33% of the students have scored between 90.1 and 95.0.

## Stated Limits and Real Limits of a Class Interval

It is relevant when working with continuous variables to define both stated and real class limits. The lower and upper stated limits, also known as apparent limits of a class, are the lowest and highest scores that could fall into that class. For example, for the class 75.1–80.0, the lower stated limit is 75.1 and the upper stated limit is 80.0.

The lower real limit is defined as the point that is midway between the stated lower limit of a class and the stated upper limit of the next lower class. The upper real limit is defined as the point that is midway between the stated upper limit of a class

and the stated lower limit of the next higher class. For example, the lower real limit of the class 80.1–85.0 is 80.05, and the upper real limit is 85.05.

Real limits may be determined not only for classes, but also for numbers. In the case of numbers, real limits are the points midway between a particular number and the next lower and higher numbers on the scale used in the respective research. For example, the lower real limit of number 4 on a 1-unit scale is 3.5, and its upper real limit is 4.5. However, real limits are not always calculated as midpoints. For example, most individuals identify their age using their most recent birthday. Thus, it is considered that a person 39 years old is at least 39 years old and has not reached his 40th birthday, and not that he is older than 38 years and 6 months and younger than 39 years and 6 months.

For discrete numbers, there are no such things as stated and real limits. When counting the number of people present at a meeting, limits do not extend below and above the respective number reported. If there are 120 people, all limits are equal to 120.

### Midpoints of Class Intervals

Each class has a midpoint defined as the point midway between the real limits of the class. Midpoints are calculated by adding the values of the stated or real limits of a class and dividing the sum by two. For example, the midpoint for the class 80.1–85.0 is $m = \frac{80.1 + 85.0}{2} = \frac{80.05 + 85.05}{2} = 82.55$.

### Advantages and Drawbacks of Using Frequency Tables

The main advantage of using frequency tables is that data are grouped and thus easier to read. Frequency tables allow the reader to immediately notice a series of characteristics of the analyzed data set that could probably not have been easily seen when looking at the raw data: the lowest score (i.e., 0, in Table 3); the highest score (i.e., 5, in Table 3); the most frequently occurring score (i.e., 5, in Table 3); and how many observations fall between two given scores (i.e., five families have between two and four children, in Table 3). Frequency tables also represent the first step in drawing histograms and calculating means from grouped data.

Using relative frequency distributions or percentage frequency tables is important when comparing the frequency distributions of samples with different sample sizes. Whereas simple frequencies depend on the total number of observations, relative frequencies and percentage frequencies do not and thus may be used for comparisons.

The main drawback of using frequency tables is the loss of detailed information. Especially when data are grouped into classes, the information for individual cases is no longer available. This means that all scores in a class are dealt with as if they were identical. For example, the reader of Table 5 learns that six students have scored between 85.1 and 90.0, but the reader does not learn any more details about the individual test results.

*Oana Pusa Mihaescu*

***See also*** Cumulative Frequency Distribution; Descriptive Statistics; Distribution; Frequency Distribution; Histogram

#### Further Readings

Downie, N. M., & Heath, R. W. (1974). *Basic statistical methods*. New York: Harper & Row.
Fielding, J. L., & Gilbert, G. N. (2000). *Understanding social statistics*. Thousand Oaks, CA: Sage.
Fried, R. (1969). *Introduction to statistics*. New York: Oxford University Press.
Hamilton, L. (1996). *Data analysis for social sciences: A first course in applied statistics*. Belmont, CA: Wadsworth.
Kiess, H. O. (2002). *Statistical concepts for the behavioral sciences*. Boston: Allyn & Bacon.
Kolstoe, R. H. (1973). *Introduction to statistics for the behavioral sciences*. Homewood, IL: Dorsey.
Lindquist, E. F. (1942). *A first course in statistics: Their use and interpretation in education and psychology*. Cambridge, MA: Riverside Press.

## Friedman Test

In an attempt to control for unwanted variability, researchers often implement designs that pair or group participants into subsets based on common

characteristics (e.g., randomized block design) or implement designs that observe the same participant across a series of conditions (e.g., repeated-measures design). The analysis of variance (ANOVA) is a common statistical method used to analyze data from a randomized block or repeated-measures design. However, the assumption of normality that underlies ANOVA is often violated, or the scale of measurement for the dependent variable is ordinal-level, hindering the use of ANOVA. To address this situation, economist Milton Friedman developed a statistical test based on ranks that may be applied to data from randomized block or repeated measures designs where the purpose is to detect differences across two or more conditions. This entry describes this statistical test, named the Friedman Test, which may be used in lieu of ANOVA. The Friedman test is classified as a nonparametric test because it does not require a specific distributional assumption. A primary advantage of the Friedman test is that it can be applied more widely as compared to ANOVA.

## Procedure

The Friedman test is used to analyze several related (i.e., dependent) samples. Friedman referred to his procedure as the *method of ranks* in that it is based on replacing the original scores with rank-ordered values. Consider a study in which data are collected within a randomized block design where $N$ blocks are observed over $K$ treatment conditions on a dependent measure that is at least ordinal-level. The first step in the Friedman test is to replace the original scores with ranks, denoted $R_{jk}$, within each block; that is, the scores for block $j$ are compared with each other, and a rank of 1 is assigned to the smallest observed score, a rank of 2 is assigned to the second smallest, and so on until the largest value is replaced by a rank of $K$. In the situation where there are ties within a block (i.e., two or more of the values are identical), the midrank is used. The midrank is the average of the ranks that would have been assigned if there were no ties. Note that this procedure generalizes to a repeated measures design in that the ranks are based on within-participant observations (or, one can think of the participants as defining the blocks). Table 1 presents the ranked data in tabular form.

**Table 1** Ranks for Randomized Block Design

| | | *Treatment Conditions* | | | | |
|---|---|---|---|---|---|---|
| | | *1* | *2* | ... | *K* | *Row Means* |
| Blocks | 1 | $R_{11}$ | $R_{12}$ | ... | $R_{1K}$ | $\overline{R}_1 = \frac{K+1}{2}$ |
| | 2 | $R_{21}$ | $R_{22}$ | ... | $R_{2K}$ | $\overline{R}_2 = \frac{K+1}{2}$ |
| | ⋮ | ... | ... | ... | ... | ⋮ |
| | $N$ | $R_{N1}$ | $R_{N2}$ | ... | $R_{NK}$ | $\overline{R}_N = \frac{K+1}{2}$ |
| Column Means | | $\overline{R}_1$ | $\overline{R}_2$ | ... | $\overline{R}_3$ | $\overline{R} = \frac{K+1}{2}$ |

It is apparent that row means in Table 1 (i.e., mean of ranks for each block) are the same across blocks; however, the column means (i.e., mean of ranks within a treatment condition) will be affected by differences across treatment conditions. Under the null hypothesis that there is no difference due to treatment, the ranks are assigned at random, and thus, an equal frequency of ranks would be expected for each treatment condition. Therefore, if there is no treatment effect, then the column means are expected to be the same for each treatment condition. The null hypothesis may be specified as follows:

$$H_0 : \mu_{\overline{R}_{\cdot 1}} = \mu_{\overline{R}_{\cdot 2}} = \cdots = \mu_{\overline{R}_{\cdot K}} = \frac{K+1}{2}. \quad (1)$$

To test the null hypothesis that there is no treatment effect, the following test statistic may be computed:

$$(TS) = \frac{N \sum_{k=1}^{K} \left(\overline{R}_{\cdot k} - \overline{R}_{\cdot\cdot}\right)^2}{\sum_{j=1}^{N} \sum_{k=1}^{K} \left(R_{jk} - \overline{R}_{\cdot\cdot}\right)^2 \Big/ (N(K-1))} \quad (2)$$

where $\overline{R}_{\cdot k}$ represents the mean value for treatment $k$; $\overline{R}_{\cdot\cdot}$ represents the grand mean (i.e., mean of all rank values); and $R_{jk}$ represents the rank for block $j$ and treatment $k$. Interestingly, the numerator and denominator of $(TS)$ can be obtained using repeated measures ANOVA on the ranks. The numerator is the sum of squares for the treatment effect ($SS_{\text{effect}}$). The denominator is the sum of squares total (which equals the sum of squares

within-subjects because there is no between-subjects variability) divided by the degrees of freedom for the treatment effect plus the degrees of freedom for the error term. Furthermore, the test statistic provided in Equation 2 does not need to be adjusted when ties exist.

An exact distribution for the test statistic may be obtained using permutation in which all possible values of ($TS$) are computed by distributing the rank values within and across blocks in all possible combinations. For an exact distribution, the $p$ value is determined by the proportion of values of ($TS$) in the exact distribution that are greater than the observed ($TS$) value. In the recent past, the use of the exact distribution in obtaining the $p$ value was not feasible due to the immense computing power required to implement the permutation. However, modern-day computers can easily construct the exact distribution for even a moderately large number of blocks. Nonetheless, for a sufficient number of blocks, the test statistic is distributed as a chi-square with degrees of freedom equal to the number of treatment conditions minus 1 (i.e., $K - 1$). Therefore, the chi-square distribution may be used to obtain the $p$ value for ($TS$) when the number of blocks is sufficient.

The Friedman test may be viewed as an extension of the sign test. In fact, in the context of two treatment conditions, the Friedman test provides the same result as the sign test. As a result, multiple comparisons may be conducted either by using the sign test or by implementing the procedure for the Friedman test on the two treatment conditions of interest. The familywise error rate can be controlled using typical methods such as Dunn-Bonferroni or Holm's Sequential Rejective Procedure. For example, when the degrees of freedom equals 2 (i.e., $K = 3$), then the Fisher least significant difference (LSD) procedure may be implemented in which the omnibus hypothesis is tested first; if the omnibus hypothesis is rejected, then each multiple comparison may be conducted using either the sign or the Friedman test on the specific treatment conditions using a full $\alpha$ level.

## Example

Suppose a researcher was interested in examining the effect of three types of exercises (weightlifting, bicycling, and running) on resting heart rate as measured by beats per minute. The researcher implemented a randomized block design in which initial resting heart rate and body weight (variables that are considered important for response to exercise) were used to assign participants into relevant blocks. Participants within each block were randomly assigned to one of the three exercise modes (i.e., treatment condition). After one month of exercising, the resting heart of each participant was recorded and is shown in Table 2.

**Table 2** Resting Heart Rate as Measured by Beats per Minute

| | | *Weight-Lifting* | *Bicycling* | *Running* |
|---|---|---|---|---|
| Block | 1 | 72 | 65 | 66 |
| | 2 | 65 | 67 | 67 |
| | 3 | 69 | 65 | 68 |
| | 4 | 65 | 61 | 60 |
| | 5 | 71 | 62 | 63 |
| | 6 | 65 | 60 | 61 |
| | 7 | 82 | 72 | 73 |
| | 8 | 83 | 71 | 70 |
| | 9 | 77 | 73 | 72 |
| | 10 | 78 | 74 | 73 |

The first step in the Friedman test is to replace the original scores with ranks within each block. For example, for the first block, the smallest original score of 65, which was associated with the participant in the bicycling group, was replaced by a rank of 1; the original score of 66 associated with the running group was replaced by a rank of 2; and the original score of 72, associated with weightlifting, was replaced by a rank of 3. Furthermore, note that for Block 2, the original values of resting heart rate were the same for the bicycling and running conditions (i.e., beats per minute equaled 67 for both conditions as shown in Table 2). Therefore, the midrank value of 2.5 was used, which was based on the average of the ranks they would have received if they were not tied (i.e., $[2 + 3]/2 = 2.5$). Table 3 reports the rank values for each block.

The mean of the ranked values for each block ($\overline{R}_{j\cdot}$) is identical because the ranks were based on within blocks. Therefore, there is no variability across blocks once the original scores have been replaced by the ranks. However, the mean of the ranks varies across treatment conditions ($\overline{R}_{\cdot k}$). If the treatment conditions are identical in the

**Table 3** Rank Values of Resting Heart Rate Within Each Block

| | | *Weight-Lifting* | *Bicycling* | *Running* | $\overline{R}_j$ |
|---|---|---|---|---|---|
| | 1 | 3 | 1 | 2 | 2 |
| | 2 | 1 | 2.5 | 2.5 | 2 |
| | 3 | 3 | 1 | 2 | 2 |
| | 4 | 3 | 2 | 1 | 2 |
| Block | 5 | 3 | 1 | 2 | 2 |
| | 6 | 3 | 1 | 2 | 2 |
| | 7 | 3 | 1 | 2 | 2 |
| | 8 | 3 | 2 | 1 | 2 |
| | 9 | 3 | 2 | 1 | 2 |
| | 10 | 3 | 2 | 1 | 2 |
| $\overline{R}_k$ | | 2.8 | 1.55 | 1.65 | 2 |

population, $\overline{R}_{\cdot k}$s are expected to be similar across the three conditions (i.e., $\overline{R}_{\cdot k} = 2$).

The omnibus test statistic, ($TS$), is computed for the data shown in Table 3 as follows:

$$(TS) = \frac{10[(2.80-2)^2 + (1.55-2)^2 + (1.65-2)^2]}{[(3-2)^2 + (1-2)^2 + \cdots + (1-2)^2]/(10(3-1))} = \frac{9.65}{19.5/20} = 9.897 \quad (3)$$

The numerator of ($TS$) is the sum of squares for the effect due to exercise routine on the ranks ($SS_{\text{effect}} = 9.65$). The denominator is the sum of squares total ($SS_{\text{total}} = 19.5$) divided by the degrees of freedom for the treatment effect ($df_{\text{effect}} = 2$) plus the degrees of freedom for the error term ($df_{\text{error}} = 18$).

The $p$ value for ($TS$) based on the exact distribution is 0.005, which allows the researcher to reject $H_0$ for $\alpha = 0.05$ and conclude that the exercise routines had a differential effect on resting heart rate. The chi-square distribution may also be used to obtain an approximated $p$ value. The omnibus test statistic follows the chi-square distribution with 2 degrees of freedom (i.e., $df = 3 - 1$) and has a $p$ value of 0.007.

To determine which methods differ, pairwise comparisons could be conducted using either the sign test or the Friedman procedure on the treatment conditions. Because $df = 2$, the Fisher LSD procedure can be used to control the familywise error rate. The omnibus test was significant at $\alpha = 0.05$, therefore, each of the pairwise comparisons can be tested using $\alpha = 0.05$. Table 4 reports the exact $p$ values (two-tailed) for the three pairwise comparisons. From the analyses, the researcher can conclude that the weightlifting condition differed in its effect on resting heart rate compared to running and bicycling; however, it cannot be concluded that the running and bicycling conditions differed.

**Table 4** $p$ Values for the Pairwise Comparisons

| *Comparison* | *p value (Exact, Two-tailed)* |
|---|---|
| Weight-lifting vs. Bicycling | 0.021 |
| Weight-lifting vs. Running | 0.021 |
| Bicycling vs. Running | 1.000 |

*Craig Stephen Wells*

***See also*** Levels of Measurement; Normal Distribution; Repeated Measures Design; Sign Test; Wilcoxon Rank Sum Test

**Further Readings**

Friedman, M. (1937a). A comparison of alternative tests of significance for the problem of $m$ rankings. *Annals of Mathematical Statistics, 11,* 86–92.

Friedman, M. (1937b). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association, 32,* 675–701.

Friedman, M. (1939). A correction: The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association, 34,* 109.

# *F* Test

The $F$ test was named by George W. Snedecor in honor of its developer, Sir Ronald A. Fisher. $F$ tests are used for a number of purposes, including comparison of two variances, analysis of variance (ANOVA), and multiple regression. An $F$ statistic is a statistic having an $F$ distribution.

## The *F* Distribution

The *F* distribution is related to the *chi-square* ($\chi^2$) distribution.

A random variable has a chi-square distribution with $m$ degrees of freedom (d.f.) if it is distributed as the sum of squares of $m$ independent standard normal random variables. The chi-square distribution arises in analysis of variance and regression analysis because the sum of squared deviations (numerator of the variance) of the dependent variable is decomposed into parts having chi-square distributions under the standard assumptions on the errors in the model.

The *F* distribution arises because one takes ratios of the various terms in the decomposition of the overall sum of squared deviations. When the errors in the model are normally distributed, these terms have distributions related to chi-square distributions, and the relevant ratios have *F* distributions. Mathematically, the *F* distribution with $m$ and $n$ d.f. is the distribution of

$$\frac{U/m}{V/n},$$

where $U$ and $V$ are statistically independent and distributed according to chi-square distributions with $m$ and $n$ d.f.

## *F* Tests

Given a null hypothesis $H_0$ and a significance level $\alpha$, the corresponding *F* test rejects $H_0$ if the value of the *F* statistic is large; more precisely, if $F > F_{\mathrm{m,n};\alpha}$, the upper $\alpha$th quantile of the $F_{m,n}$ distribution. The values of m and n depend upon the particular problem (comparing variances, ANOVA, multiple regression). The achieved (descriptive) level of significance ($p$ value) of the test is the probability that a variable with the $F_{m,n}$ distribution exceeds the observed value of the statistic *F*. The null hypothesis is rejected if $p < \alpha$.

Many tables are available for the quantiles, but they can be obtained in Excel and in statistical computer packages, and $p$ values are given in the output for various procedures.

## Comparing Two Variances

One sort of *F* test is that for comparing two independent variances. The sample variances are compared by taking their ratio. This problem is mentioned first, as other applications of *F* tests involve ratios of variances (or mean squares) as well. The hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ is rejected if the ratio

$$F = \frac{\text{larger sample variance}}{\text{smaller sample variance}}$$

is large. The statistic has an *F* distribution if the samples are from normal distributions.

## Normal Distribution Theory

The distribution of the sample variance $s^2$ computed from a sample of $N$ from a normal distribution with variance $\sigma^2$ is given by the fact that $(N-1)s^2/\sigma^2$ is distributed according to a chi-square distribution with $m = N - 1$ d.f. So, for the variances $s_1^2$ and $s_2^2$ of two independent samples of sizes $N_1$ and $N_2$ from normal distributions, the variable $U = (N_1 - 1)s_1^2/\sigma_1^2$ is distributed according to chi-square with $m = N_1 - 1$ d.f., and the variable $V = (N_2 - 1)s_2^2/\sigma_2^2$ is distributed according to chi-square with $n = N_2 - 1$ d.f. If $\sigma_1^2 = \sigma_2^2$, the ratio $F = s_1^2/s_2^2$ has an *F* distribution with $m = N_1 - 1$ and $n = N_2 - 1$ d.f.

## Analysis of Variance

*F* tests are used in ANOVA. The total sum of squared deviations is decomposed into parts corresponding to different factors. In the normal case, these parts have distributions related to chi-square. The *F* statistics are ratios of these parts and hence have *F* distributions in the normal case.

### *One-Way ANOVA*

One-way ANOVA is for comparison of the means of several groups. The data are $Y_{gi,g} = 1, 2, \ldots, k$ groups, $i = 1, 2, \ldots, N_g$ cases in the $g$th group.

The model is

$$Y_{gi} = \mu_g + \varepsilon_{gi},$$

where the errors $\varepsilon_{gi}$ have mean zero and constant variance $\sigma^2$, and are uncorrelated. The null hypothesis (hypothesis of no differences between group means) is $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$. It is convenient to reparametrize as $\mu_g = \mu + \alpha_g$, where $\alpha_g$ is the deviation of the true mean $\mu_g$ for group $g$ from the true overall mean $\mu$ The deviations satisfy a constraint such as $\sum_{g=1}^{k} N_g \alpha_g = 0$. In terms of the $\alpha_g$, the model is

$$Y_{gi} = \mu + \alpha_g + \varepsilon_{gi},$$

and $H_0$ is $\alpha_1 = \alpha_2 = \cdots = \alpha_k$.

### *Decomposition of Sum of Squares; Mean Squares*

There is a corresponding decomposition of the observations and of the sums of squared deviations from the mean. The group sums are $Y_{g+} = \sum_{i=1}^{N_g} Y_{gi}$. The means are $Y_{g.} = Y_{g}+/N_g$. The overall sum is $Y_{++} = \sum_{g=1}^{k} \sum_{i=1}^{N_g} Y_{gi}$, and the overall mean is $Y.. = Y_{++}/N$, where $N = N_1 + N_2 + \cdots + N_K$. The decomposition of the observations as

$$\begin{aligned} Y_{gi} &= \hat{\mu} + \hat{\alpha}_g + \hat{\varepsilon}_i \\ &= Y.. + (Y_{g.} - Y_{++}) + (Y_{gi} - Y_{g.}). \end{aligned}$$

This is

$$(Y_{gi} - Y..) = (Y_{g.} - Y.. + (Y_{gi} - Y_{g.}).$$

Squaring both sides and summing gives the analogous decomposition of the sum of squares

$$\sum_{g=1}^{k} \sum_{i=1}^{N_g} (Y_{gi} - Y..)^2 =$$

$$\sum_{g=1}^{k} \sum_{i=1}^{N_g} (Y_{g.} - Y..)^2 + \sum_{g=1}^{k} \sum_{i=1}^{N_g} (Y_{gi} - Y_{g.})^2 =$$

$$\sum_{g=1}^{k} N_g (Y_{g.} - Y..)^2 + \sum_{g=1}^{k} \sum_{i=1}^{N_g} (Y_{gi} - Y_{g.})^2$$

or $SSTot = SSB + SSW$.

Here, *SSTot* denotes the total sum of squares; *SSB*, between-group sum of squares; and *SSW*, within-group sum of squares. The decomposition of d.f. is

$$DFTot = DFB + DFW;$$

that is,

$$(N - 1) = (k - 1) + (N - k).$$

Each mean square is the corresponding sum of squares, divided by its d.f.: $MSTot = SSTot/DFTot = SSTot/(N - 1)$ is just the sample variance of *Y*; $MSB = SSB/DFB = SSB/(k - 1)$, and $MSW = SSW/DFW = SSW/(N - k)$. The relevant *F* statistic is $F = MSB/MSW$. For *F* to have an *F* distribution, the errors must be normally distributed. The residuals can be examined to see if their histogram looks bell-shaped and not too heavy-tailed, and a normal quantile plot can be used.

### *Power and Noncentral F*

The power of the test depends upon the extent of departure from the null hypothesis, as given by the noncentrality parameter $\delta^2$. For one-way ANOVA,

$$\sigma^2 \delta^2 = \sum_{g=1}^{k} N_g (\mu_g - \mu)^2 = \sum_{g=1}^{k} N_g \alpha_g^2,$$

a measure of dispersion of the true means $\mu_g$. The noncentral *F* distribution is related to the noncentral chi-square distribution. The noncentral chi-square distribution with $m$ degrees of freedom and noncentrality parameter $\delta^2$ is the distribution of the sum of squares of $m$ independent normal variables with variances equal to 1 and means whose sum of squares is $\delta^2$. If, in the ratio $(U/m)/(V/n)$, the variable $U$ has a noncentral chi-square distribution, then the ratio has a noncentral *F* distribution. When the null hypothesis of equality of means is false, the test statistic has a noncentral *F* distribution. The noncentrality parameter depends upon the group means and the sample sizes. Power computations involve the noncentral *F* distribution. It is via the noncentrality parameter that one specifies what constitutes a reasonably large departure from the null hypothesis. Ideally, the level $\alpha$ and the sample sizes are chosen so that the power is sufficiently large (say, .8 or .9) for large departures from the null hypothesis.

### *Randomized Blocks Design*

This is two-way ANOVA with no replication. There are two factors *A* and *B* with $a$ and $b$ levels,

thus, $ab$ observations. The decomposition of the sum of squares is

$$SSTot = SSA + SSB + SSRes.$$

The decomposition of d.f. is

$$DFTot = DFA + DFB + DFRes$$

that is,

$$(ab - 1) = (a - 1) + (b - 1) + (a - 1)(b - 1).$$

Each mean square is the corresponding sum of squares, divided by its d.f.: $MSA = SSA/DFA = SSA/(a - 1)$, $MSB = SSB/DFB = SSB/(b - 1)$, $MSRes = SSRes/DFRes = SSRes/(a - 1)(b - 1)$. The test statistics are $F_A = MSA/MSRes$ with $a - 1$ and $(a - 1)(b - 1)$ d.f. and $F_B = MSB/MSRes$ with $b - 1$ and $(a - 1)(b - 1)$ d.f.

### *Two-Way ANOVA*

When there is replication, with $r$ replicates for each combination of levels of $A$ and $B$, the decomposition of $SSTot$ is

$$SSTot = SSA + SSB + SSA \times B + SSRes,$$

where $A \times B$ represents interaction. The decomposition of d.f. is

$$DFTot = DFA + DFB + DFA \times B + DFRes$$

that is,

$$(abr - 1) = (a - 1) + (b - 1) + (a - 1)(b - 1) + ab(r - 1).$$

Each mean square is the corresponding sum of squares, divided by its d.f.: $MSA = SSA/DFA = SSA/(a - 1)$, $MSB = SSB/DFB = SSB/(b - 1)$, $MSA \times B = SSA \times B/DFA \times B = SSA \times B/(a - 1)(b - 1)$, $MSRes = SSRes/DFRes = SSRes/ab(r - 1)$. The denominators of the $F$ tests depend on whether the levels of the factors are fixed, random, or mixed.

### *Other Designs*

The concepts developed above for one-way and two-way ANOVA should enable the reader to understand the $F$ tests for more complicated designs.

## Multiple Regression

Given a data set of observations on explanatory variables $X_1, X_2, \ldots, X_p$ and a dependent variable $Y$ for each of $N$ cases, the multiple linear regression model takes the expected value of $Y$ to be a linear function of $X_1, X_2, \ldots, X_p$. That is, the mean $E_x(Y)$ of Y for given values $x_1, x_2, \ldots, x_p$ is of the form

$$E_x(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p,$$

where the $\beta_j$ are parameters to be estimated, and the error is additive,

$$Y = E_x(Y) + \varepsilon.$$

Writing this in terms of the $N$ cases gives the observational model for $i = 1, 2, \ldots, N$,

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \varepsilon_i.$$

The assumptions on the errors are that they have mean zero and common variance $\sigma^2$, and are uncorrelated.

### *Decomposition of Sum of Squares*

Analogously to the model $Y_i = E_x(Y_i) + \varepsilon_i$, the observations can be written as

$$\text{Observed value} = \text{predicted value} + \text{residual},$$

or

$$Y_i = \hat{Y}_i + e_i.$$

Here, $\hat{Y}_i$ is the predicted value of $Y_i$, given by

$$\hat{Y}_i = b_0 + b_{1x1i} + b_{2x2i} + \cdots + b_{pxpi},$$

where $b_j$ is the least squares estimate of $\beta_j$. This can be written

$$\hat{Y}_i = Y. + b_1(x_{1i} - x_{1.}) + b_2(x_{2i} - x_{2.}) + \ldots b_p(x_{pi} - x_{p.}).$$

Here, $Y.$ is the mean of the values of $Y$ and $x_{j.}$ is the mean of the values of $X_j$. This gives

$$(Y_i - Y.) = (\hat{Y}_i - Y.) + (Y_i - \hat{Y}_i),$$

where the residual $e_i = Y_i - \hat{Y}_i$ estimates the error $\varepsilon_i$. Squaring and summing gives the decomposition of sum of squares

$$SSTot = SSReg + SSRes.$$

Here *SSReg* is the regression sum of squares and *SSRes* is the residual sum of squares. The proportion of *SSTot* accounted for by the regression is *SSReg/SSTot* = *SSReg*/(*SSReg* + *SSRes*), which can be shown to be equal to $R^2$, where $R$ is the multiple correlation coefficient between $Y$ and the $p$ $X$s.

### *F* in Terms of R-square

The decomposition

$$SSTot = SSReg + SSRes$$

is $SSTot = R2SSTot + (1 - R^2)SSTot$. The decomposition of d.f. is

$$DFTot = DFReg + DFRes,$$

or

$$(N - 1) = p + (N - p - 1).$$

It follows that $MSReg = SSReg/DFReg = SSReg/p = R^2/p$ and $MSRes = SSRes/DFRes = SSRes/(n - p - 1) = (1 - R^2)/(N - p - 1)$. The statistic

$$F = MSReg/MSRes = \frac{R^2/p}{(1 - R^2)/(N - p - 1)}$$

has an $F$ distribution with $p$ and $N - p - 1$ d.f. when the errors are normally distributed.

### *Comparing a Reduced Model With the Full Model*

The $F$ statistic for testing the null hypothesis concerning the coefficient of a single variable is the square of the $t$ statistic for this test. But $F$ can be used for testing several variables at a time. It is often of interest to test a portion of a model, that is, to test whether a subset of the variables—say, the first $q$ variables—is adequate. Let $p = q + r$; it is being considered whether the last $r = p - q$ variables are needed. The null hypothesis is

$$H_0 : \beta_j = 0, j = q + 1, \ldots, p.$$

When this null hypothesis is true, the model is the reduced model

$$Ex(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q.$$

An $F$ test can be used to test the adequacy of this reduced model. Let $SSRes_{full}$ denote the residual sum of squares of the full model with $p$ variables and $SSRes_{red}$ that of the reduced model with only $q$ variables. Then the test statistic is

$$F = \frac{(SSRes_{red} - SSRes_{full})/(p - q)}{SSRes_{full}/(N - p - 1)}.$$

The denominator $SSRes_{full}/(N - p - 1)$ is simply $MSRes_{full}$, the residual mean square in the full model. The quantity $SSRes_{full}$ is less than or equal to $SSRes_{red}$ because the latter is the result of minimization over a subset of the regression coefficients, the last $p - q$ of them being restricted to zero. The difference $SSRes_{red} - SSRes_{full}$ is thus nonnegative and is the loss of fit due to omitting the last $p - q$ variables, that is, the loss of fit due to the null hypothesis that the last $p - q$ $\beta_j$ are zero. It is the hypothesis sum of squares *SSHyp*. Thus, the numerator is

$$(SSRes_{red} - SSRes_{full})/(p - q) = SSHyp/(p - q) = MSHyp,$$

the hypothesis mean square. So $F$ is again the ratio of two mean squares, in this case, $F = MSHyp/MSRes_{full}$. The numbers of d.f. are $p - q$ and $N - p - 1$. Under the null hypothesis, this $F$ statistic has the $F_{p-q,\, N-p-1}$ distribution if the errors are normally distributed.

*Stanley L. Sclove*

***See also*** Analysis of Variance (ANOVA); Coefficients of Correlation, Alienation, and Determination; Experimental Design; Factorial Design; Hypothesis; Least Squares, Methods of; Significance Level, Concept of; Significance Level, Interpretation and Construction; Stepwise Regression

### Further Readings

Bennett, J. H. (Ed.). (1971). *The collected papers of R. A. Fisher.* Adelaide, Australia: University of Adelaide Press.

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh, UK: Oliver and Boyd.
Graybill, F. A. (1976). *Theory and application of the linear model*. N. Scituate, MA: Duxbury Press.
Hogg, R. V., McKean, J. W., & Craig, A. T. (2005). *Introduction to mathematical statistics* (6th ed.). Upper Saddle River, NJ: Prentice Hall.
Kempthorne, O. (1952). *The design and analysis of experiments*. New York: Wiley.
Scheffé, H. (1959). *The analysis of variance*. New York: Wiley.
Snedecor, G. W., & Cochran, W. G. (1989). *Statistical methods* (8th ed.). Ames: Iowa State University Press.

# G

## Gain Scores, Analysis of

Gain (i.e., change, difference) is defined here as the difference between test scores obtained for an individual or group of individuals from a measurement instrument, intended to measure the same attribute, trait, concept, construct, or skill, between two or more testing occasions. This difference does not necessarily mean that there is an increase in the test score(s). Thus, a negative difference is also described as a "gain score." There are a multitude of reasons for measuring gain: (a) to evaluate the effects of instruction or other treatments over time, (b) to find variables that correlate with change for developing a criterion variable in an attempt to answer questions such as "What kinds of students grow fastest on the trait of interest?," and (c) to compare individual differences in gain scores for the purpose of allocating service resources and selecting individuals for further or special study.

The typical and most intuitive approach to the calculation of change is to compute the difference between two measurement occasions. This difference is called a *gain score* and can be considered a composite in that it is made up of a pretest (e.g., an initial score on some trait) and a posttest (e.g., a score on the same trait after a treatment has been implemented) score where a weight of 1 is assigned to the posttest and a weight of $-1$ is assigned to the pretest. Therefore, the computation of the gain score is simply the difference between posttest and pretest scores (i.e., gain = posttest − pretest). However, both the pretest and the posttest scores for any individual contain some amount of measurement error such that it is impossible to know a person's true score on any given assessment. Thus, in classical test theory (CTT), a person's observed score ($X$) is composed of two parts, some true score ($T$) and some amount of measurement error ($E$) as defined in Equation 1:

$$X = T + E. \tag{1}$$

In a gain score analysis, it is the change in the true scores ($\Delta_T$) that is of real interest. However, the researcher's best estimate of the true score is the person's observed score, thus making the gain score (i.e., the difference between observed scores) an unbiased estimator of $\Delta_T$ for any given individual or subject. What follows is a description of methods for analyzing gain scores, a discussion of the reliability of gain scores, alternatives to the analysis of gain scores, and a brief overview of designs that measure change using more than two waves of data collection.

### Methods for the Analysis of Gain Scores

The gain score can be used as a dependent variable in a *t* test (i.e., used to determine whether the mean difference is statistically significant for a group or whether the mean differences between two groups are statistically significantly different) or an analysis of variance (ANOVA) (i.e., used when the means of more than two groups or more

than two measurement occasions are compared) with the treatment, intervention, instructional mode (i.e., as with educational research) or naturally occurring group (e.g., sex) serving as the between-subjects factor. (For simplicity, throughout this entry, levels of the between-groups factors are referred to as treatment groups. However, the information provided also applies to other types of groups as well, such as intervention, instructional modes, and naturally occurring.) If the *t* test or the treatment main effect in an ANOVA is significant, the null hypothesis of no significant gain or difference in improvement between groups (e.g., treatment and control groups) can be rejected.

### t *Test*

Depending on the research question and design, either a one-sample *t* test or an independent samples *t* test can be conducted using the gain score as the dependent variable. A one-sample *t* test can be used when the goal is to determine whether the mean gain score is significantly different from zero or some other specified value. When two groups (e.g., control and treatment) are included in the research design and the aim is to determine whether more gain is observed in the treatment group, for example, an independent *t* test can be implemented to determine whether the mean gain scores between groups are significantly different from each other. In this context, the gain score is entered as the dependent variable and more than two groups would be examined (e.g., a control and two different treatment groups).

### *Analysis of Variance*

Like the goal of an independent *t* test, the aim of an ANOVA is to determine whether the mean gain scores between groups are significantly different from each other. Instead of conducting multiple *t* tests, an ANOVA is performed when more than two groups are present in order to control the type I error rate (i.e., rate of rejecting a true null hypothesis). However, differences in pretest scores between groups are not controlled for when conducting an ANOVA using the gain scores, which can result in misleading conclusions as discussed later.

## Reliability of the Gain Score

Frederic M. Lord and Melvin R. Novick introduced the reliability of the gain score as the ratio of the variance of the difference score ($\sigma_D^2$) to the sum of the variance of the difference score and the variance of the error associated with that difference ($\sigma_D^2 + \sigma_{err_D}^2$):

$$\rho_D = \frac{\sigma_D^2}{\sigma_D^2 + \sigma_{err_D}^2}. \quad (2)$$

Hence, the variance of the difference score is the systematic difference between subjects in their gain score. In other words, the reliability of the gain score is really a way to determine whether the assessment or treatment discriminates between those who change a great deal and those who change little, and to what degree. The reliability of the gain score can be further described in terms of the pretest and posttest variances along with their respective reliabilities and the correlation of the pretest with the posttest. Equation 3 describes this relationship from the CTT perspective, where observations are considered independent.

$$\rho_D = \frac{\sigma_{pre}^2\rho_{pre} + \sigma_{post}^2\rho_{post} - 2\sigma_{pre}\sigma_{post}\rho_{pre,post}}{\sigma_{pre}^2 + \sigma_{post}^2 - 2\sigma_{pre}\sigma_{post}\rho_{pre,post}}, \quad (3)$$

where $\rho_D$ represents the reliability of the gain score ($D$) and $\sigma_{pre}^2$ and $\sigma_{post}^2$ designate the variance of the pretest and posttest scores, respectively. Likewise, $\sigma_{pre}$ and $\sigma_{post}$ designate the standard deviations of the pretest and posttest scores, respectively, and $\rho_{pre}$ and $\rho_{post}$ represent the reliabilities of the pretest and posttest scores, respectively. Lastly, $\rho_{pre,post}$ designates the correlation between the pretest and posttest scores. Equation 3 further reduces to Equation 4:

$$\rho_D = \frac{\rho_{pre} + \rho_{post} - 2\rho_{pre,post}}{2 - 2\rho_{pre,post}} \quad (4)$$

when the variances of the pretests and posttests are equal (i.e., $\sigma_{pre}^2 = \sigma_{post}^2$). However, it is rare that equal variances are observed when a treatment is studied that is intended to show growth between pretesting and posttesting occasions. When growth is the main criterion, this equality should not be considered an indicator of construct validity, as it

has been in the past. In this case, it is merely an indication of whether rank order is maintained over time. If differing growth rates are observed, this equality will not hold. For example, effective instruction tends to increase the variability within a treatment group, especially when the measure used to assess performance has an ample number of score points to detect growth adequately (i.e., the scoring range is high enough to prevent ceiling effects). If ceiling effects are present or many students achieve mastery such that scores are concentrated near the top of the scoring scale, the variability of the scores declines.

The correlation between pretest and posttest scores for the treatment group provides an estimate of the reliability (i.e., consistency) of the treatment effect across individuals. When the correlation between the pretest and posttest is one, the reliability of the difference score is zero. This is because uniform responses are observed, and therefore, there is no ability to discriminate between those who change a great deal and those who change little. However, some researchers, Gideon J. Mellenbergh and Wulfert P. van den Brink, for example, suggest that this does not mean that the difference score should not be trusted. In this specific instance, a different measure (e.g., measure of sensitivity) is needed to assess the utility of the assessment or the productivity of the treatment in question. Such measures may include, but are not limited to, Cohen's effect size or an investigation of information (i.e., precision) at the subject level.

Additionally, experimental independence (i.e., the pretest and posttest error scores are uncorrelated) is assumed by using the CTT formulation of reliability of the difference score. This is hardly the case with educational research, and it is likely that the errors are positively correlated; thus, the reliability of gain scores is often underestimated. As a result, in cases such as these, the additivity of error variances does not hold and leads to an inflated estimate of error variance for the gain score. Additionally, David R. Rogosa and John B. Willett contend that it is not the positive correlation of errors of measurement that inflate the reliability of the gain score, but rather individual differences in true change.

Contrary to historical findings, Donald W. Zimmerman and Richard H. Williams, among others, have shown that gain scores can be reliable under certain circumstances that depend upon the experimental procedure and the use of appropriate instruments. Williams, Zimmerman, and Roy D. Mazzagatti further discovered that for simple gains to be reliable, it is necessary that the intervention or treatment be strong and the measuring device or assessment be sensitive enough to detect changes due to the intervention or treatment. The question remains, "How often does this occur in practice?" Zimmerman and Williams show, by example, that with a pretest assessment that has a 0.9 reliability, if the intervention increases the variability of true scores, the reliability of the gain scores will be at least as high as that of the pretest scores. Conversely, if the intervention reduces the variability of true scores, the reliability of the gain scores decreases, thus placing its value between the reliabilities of the pretest and posttest scores. Given these findings, it seems that the use of gain scores in research is not as meek as it was once thought. In fact, only when there is no change or a reduction in the variance of the true scores as a result of the intervention(s) is the reliability of the gain score significantly lowered. Thus, when pretest scores are reliable, gain scores are reliable for research purposes.

Although the efficacy of using gain scores has been historically wrought with much controversy, as the main arguments against their use are that they are unreliable and negatively correlated with pretest scores, gain scores are currently gaining in application and appeal because of the resolution of misconceptions found in the literature on the reliability of gain scores. Moreover, depending on the research question, precision may be a better way to judge the utility of the gain score than reliability alone.

## Alternative Analyses

Alternative statistical tests of significance can also be performed that do not include a direct analysis of the gain scores. An analysis of covariance (ANCOVA), residualized gain scores, and the Johnson-Neyman technique are examples. Many other examples also exist but are not presented here (see Further Readings for references to these alternatives).

### *Analysis of Covariance and Residualized Gain Scores*

Instead of using the actual difference between pretest and posttest scores, Lee J. Cronbach and Lita Furby suggest the use of ANCOVA or residualized gain scores. In such an analysis, a regression line is fit that relates the pretest to the posttest scores. Then, the difference between the posttest score and predicted posttest scores is used as an indication of individuals who have changed more or less than expected, assuming a linear relationship. More specifically, the gain score is transformed into a residual score ($z$), where

$$z = x_2 - [b_x(x_1) + a_x], \tag{5}$$

$x_2$ represents the posttest score, and $b_x(x_1) + a_x$ represents the linear regression estimate of $x_2$ based on $x_1$, the pretest score.

Such an analysis answers the research question, "What is the effect of the treatment on the posttest that is not predictable from the pretest?" In other words, group means on the posttest are compared conditional on the pretest scores. Usually, the power to detect differences between groups is greater using this analysis than it is using an ANOVA. Additional advantages include the ability to test assumptions of linearity and homogeneity of regression.

### *Johnson-Neyman Technique*

In the case of quasi-experimental designs, the gain score may serve to eliminate initial differences between groups (i.e., due to the non-random assignment of individuals to treatment groups). However, post hoc adjustments for differences between groups on a pretest, such as calculation of gain scores, repeated measures ANOVA, and ANCOVA, all assume that in the absence of a treatment effect, the treatment and control groups grow at the same rate. There are two well-known scenarios (e.g., fan-spread and ceiling effects) where this is not the case, and the Johnson-Neyman technique may be an appropriate alternative for inferring treatment effects.

In the case of fan-spread, the less capable examinees may have a greater chance for improvement (e.g., in the case of special education interventions) and have greater pretest scores without the intervention when compared to those with greater impairment. Analytic post hoc methods for controlling pretreatment differences would wrongly conclude an effect due to the intervention in this case. Likewise, with high pretest scores and an assessment that restricts the range of possible high scores, a ceiling effect can occur where an underestimate of the treatment's efficacy is observed. The strength of the Johnson-Neyman technique is that it can define regions along the continuum of the covariate (i.e., the pretest score) for which the conclusion of a significant difference of means on the posttest can be determined.

## Multiple Observation Designs

Up to this point, the pre/posttest (i.e., two-wave) design has been discussed, including some of its limitations. However, the real issue may be that the assumption that change can be measured adequately from using only two scores is unrealistic and unnatural. In the case where more than two observations are present, it is possible to fit a statistical curve (i.e., growth curve) to the observed data to model change as a function of time. Thus, linearity is no longer assumed because change is realistically a continuous process and not a quantum leap (i.e., incremental) from one point to the next. With more data points, the within-subject error decreases and the statistical power increases. Analytic models describing the relationship between individual growth and time have been proposed. These models are used to estimate their respective parameters using various regression analysis techniques. Thus, the reliability of estimates of individual change can be improved by having more than two waves of data, and formulas for determining reliability for these models can be found.

## Final Thoughts

The purpose for analyzing gain scores is to either examine overall effects of treatments (i.e., population change) or distinguish individual differences in response to the treatment (i.e., individual change). Reliability of the gain score is certainly more of a concern for the former purpose, whereas precision of information is of primary concern for the

estimation of individual gain scores. Methods for analyzing gain scores include, but are not limited to, *t* tests and ANOVA models. These models answer the question "What is the effect of the treatment on change from pretest to posttest?" Gain scores focus on the difference between measurements taken at two points in time and thus represent an incremental model of change. Ultimately, multiple waves of data should be considered for the analysis of individual change over time because it is unrealistic to view the process of change as following a linear and incremental pattern.

*Tia Sukin*

***See also*** Analysis of Covariance (ANCOVA); Analysis of Variance (ANOVA); Growth Curves

**Further Readings**


Cronbach, L. J., & Furby, L. (1970). How we should measure "change"—or should we? *Psychological Bulletin, 74*(1), 68–80.

Haertel, E. H. (2006). Reliability. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65–110). Westport, CT: Praeger.

Johnson, P. O., & Fay, L. C. (1950). The Johnson-Neyman technique, its theory and application. *Psychometrika, 15*(4), 349–367.

Knapp, T. R., & Schafer, W. D. (2009). From gain score t to ANCOVA F (and vice versa). *Practical Assessment, Research & Evaluation, 14*(6), 1–7.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test development*. Reading, MA: Addison-Wesley.

Mellenbergh, G. J., & van den Brink, W. P. (1998). The measurement of individual change. *Psychological Methods, 3*(4), 470–485.

Rogosa, D., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin, 92*(3), 726–748.

Rogosa, D. R., & Willett, J. B. (1985). Understanding correlates of change by modeling individual differences in growth. *Psychometrika, 50,* 203–228.

Williams, R. H., & Zimmerman, D. W. (1996). Are simple gain scores obsolete? *Applied Psychological Measurement, 20*(1), 59–69.

Williams, R. H., Zimmerman, D. W., & Mazzagatti, R. D. (1987). Large sample estimates of the reliability of simple, residualized, and base-free gain scores. *Journal of Experimental Education, 55*(2), 116–118.

Zimmerman, D. W., & Williams, R. H. (1982). Gain scores in research can be highly reliable. *Journal of Educational Measurement, 19*(2), 149–154.

Zimmerman, D. W., & Williams, R. H. (1998). Reliability of gain scores under realistic assumptions about properties of pre-test and post-test scores. *British Journal of Mathematical and Statistical Psychology, 51*(2), 343–351.


## Game Theory

Game theory is a model of decision making and strategy under differing conditions of uncertainty. Games are defined as strategic interactions between players, where strategy refers to a complete plan of action including all prospective play options as well as the player's associated outcome preferences. The formal predicted strategy for solving a game is referred to as a solution. The purpose of game theory is to explore differing solutions (i.e., tactics) among players within games of strategy that obtain a maximum of utility. In game theory parlance, "utility" refers to preferred outcomes that may vary among individual players. John von Neumann and Oskar Morgenstern seeded game theory as an economic explanatory construct for all endeavors of the individual to achieve maximum utility or, in economic terms, profit; this is referred to as a maximum. Since its inception in 1944, game theory has become an accepted multidisciplinary model for social exchange in decision making within the spheres of biology, sociology, political science, business, and psychology. The discipline of psychology has embraced applied game theory as a model for conflict resolution between couples, within families, and between hostile countries; as such, it is also referred to as the theory of social situations.

Classic game theory as proposed by von Neumann and Morgenstern is a mathematical model founded in utility theory, wherein the game player's imagined outcome preferences can be combined and weighted by their probabilities. These outcome preferences can be quantified and are therefore labeled utilities. A fundamental assumption of von Neumann and Morgenstern's game theory is that the game player or decision maker has clear preferences and expectations. Each player is presumed rational in his or her choice behavior, applying logical heuristics in weighing all choice options, thereby formulating

his or her game strategy in an attempt to optimize the outcome by solving for the maximum. These game strategies may or may not be effective in solving for the maximum; however, the reasoning in finding a solution must be sound.

## Game Context

Three important contextual qualities of a game involve whether the game is competitive or noncompetitive, the number of players involved in the game, and the degree to which all prior actions are known.

Games may be either among individuals, wherein they are referred to as competitive games, or between groups of individuals, typically characterized as noncompetitive games. The bulk of game theory focuses on competitive games of conflict. The second contextual quality of a game involves player or group number. Although there are situations in which a single decision maker must choose an optimal solution without reference to other game players (i.e., human against nature), generally, games are between two or more players. Games of two players or groups of players are referred to as two-person or two-player (where "player" may reflect a single individual or a single group of individuals) games; these kinds of games are models of social exchange. In a single-person model, the decision maker controls all variables in a given problem; the challenge in finding an optimal outcome (i.e., maximum) is in the number of variables and the nature of the function to be maximized. In contrast, in two-person, two-player or n-person, n-player games (where "n" is the actual number of persons or groups greater than two), the challenge of optimization hinges on the fact that each participant is part of a social exchange, where each player's outcome is interdependent on the actions of all other players. The variables in a social exchange economy are the weighted actions of all other game players.

The third important quality of a game is the degree to which players are aware of other players' previous actions or moves within a game. This awareness is referred to as information, and there are two kinds of game information: perfect and imperfect. Games of perfect information are those in which all players are aware of all actions during the game; in other words, if Player A moves a pawn along a chess board, Player B can track that pawn throughout the game. There are no unknown moves. This is referred to as perfect information because there is no uncertainty, and thus, games of perfect information have few conceptual problems; by and large, they are considered technical problems. In contrast, games of imperfect information involve previously unknown game plans; consequently, players are not privy to all previously employed competitive strategies. Games of imperfect information require players to use Bayesian interpretations of others' actions.

## The Role of Equilibrium

Equilibrium refers to a stable outcome of a game associated with two or more strategies and by extension, two or more players. In an equilibrium state, player solutions are balanced, the resources demanded and the resources available are equal, this means that one of the two parties will not optimize. John Forbes Nash provided a significant contribution to game theory by proposing a conceptual solution to analyze strategic interactions and consequently the strategic options for each game player, what has come to be called Nash equilibrium. This equilibrium is a static state, such that all players are solving for optimization and none of the players benefit from a unilateral strategy change. In other words, in a two-person game, if player A changes strategy and player B does not, player A has departed from optimization; the same would be true if player B changed strategy in the absence of a strategy change by player A. Nash equilibrium of a strategic game is considered stable because all players are deadlocked, their interests are evenly balanced and in the absence of some external force, like a compromise, they are unlikely to change their tactical plan. Nash equilibrium among differing strategic games has become a heavily published area of inquiry within game theory.

## Types of Games

Games of strategy are typically categorized as zero-sum games (also known as constant-sum

games), non–zero-sum competitive games, and non–zero-sum cooperative games; within this latter category are also bargaining games and coalitional games.

### *Zero-Sum Games*

A defining feature of zero-sum games is that they are inherently win-lose games. Games of strategy are characterized as zero-sum or constant-sum games if the additive gain of all players is equal to zero. Two examples of zero-sum games are a coin toss or a game of chicken. Coin tosses are strictly competitive zero-sum games, where a player calls the coin while it is still aloft. The probability of a head or a tail is exactly 50:50. In a coin toss, there is an absence of Nash equilibrium; given that there is no way to anticipate accurately what the opposing player will choose, nor is it possible to predict the outcome of the toss, there exists only one strategic option—choose. Consequently, the payoff matrix in a coin toss contains only two variables: win or lose. If Player A called the toss inaccurately, then his or her net gain is −1, and player B's gain was +1. There is no draw. Player A is the clear loser, and Player B is the clear winner. However, not all zero-sum outcomes are mutually exclusive; draws can occur, for example, in a two-player vehicular game of chicken, where there are two car drivers racing toward each other. The goal for both drivers is to avoid yielding to the other driver; the first to swerve away from the impending collision has lost the game. In this game, there are four possible outcomes: Driver A yields, Driver B yields, neither Driver A nor B yields, or both Driver A and Driver B simultaneously swerve. However, for Driver A, there are only two strategic options: Optimize his or her outcome (+1) or optimize the outcome for Driver B (−1); Player B has these same diametrically opposed options. Note that optimizing for the maximum is defined as winning. Yet surviving in the absence of a win is losing, and dying results in a forfeited win, so there is no Nash equilibrium in this zero-sum game, either. Table 1 reflects the outcome matrix for each driver in a game of chicken, where the numeric values represent wins (+) and losses (−).

**Table 1** Outcome Matrix for a Standard Game of Chicken

| | | *Driver B* | |
|---|---|---|---|
| | | *Yield* | *Maintain* |
| *Driver A* | *Yield* | 0, 0 | −1, +1<br>Driver B wins |
| | *Maintain* | +1, −1<br>Driver A wins | 0, 0 |

### *Non–Zero-Sum Competitive Games*

Non–zero-sum games of strategy are characterized as situations in which the additive gain of all players is either more than or less than zero. Non-zero-sum games may yield situations in which players are compelled by probability of failure to depart from their preferred strategy in favor of another strategy that does the least violence to their outcome preference. This kind of decisional strategy is referred to as a minimax approach, as the player's goal is to minimize his or her maximum loss. However, the player may also select a decisional strategy that yields a small gain. In this instance, he or she selects a strategic solution that maximizes the minimum gain; this is referred to as a maximin. In two-person, non–zero-sum games, if the maximin for one player and the minimax for another player are equal, then the two players have reached Nash equilibrium. In truly competitive zero-sum games, there can be no Nash equilibrium. However, in non–zero-sum games, Nash equilibria are frequently achieved; in fact, an analogous non–zero-sum game of zero-sum chicken results in two Nash equilibria. Biologists and animal behaviorists generally refer to this game as Hawk-Dove in reference to the aggressive strategies employed by the two differing species of birds.

Hawk-Dove games are played by a host of animal taxa, including humans. Generally, the context for Hawk-Dove involves conspecific species competing for indivisible resources such as access to mating partners, such that an animal must employ an aggressive Hawk display and attack strategy or a noncommittal aggressive display reflective of a Dove strategy. The Hawk tactic is a show of aggressive force in conjunction with a commitment

**Table 2** Outcome Matrix for the Hawk-Dove Game

| | | *Conspecific B* | |
|---|---|---|---|
| | | *Hawk* | *Dove* |
| *Conspecific A* | *Hawk* | 1, 1 | 4, 2 |
| | *Dove* | 2, 4 | 3, 3 |

**Table 3** Outcome Matrix for the Prisoner's Dilemma

| | | *Prisoner B* | |
|---|---|---|---|
| | | *Defect* | *Cooperate* |
| *Prisoner A* | *Defect* | 5, 5 | 0, 10 |
| | *Cooperate* | 10, 0 | 10, 10 |

to follow through on the aggressive display with an attack. The Dove strategy employs a display of aggressive force without a commitment to follow up the show, thereby fleeing in response to a competitive challenge. The goal for any one animal is to employ a Hawk strategy while the competitor uses a Dove strategy. Two Hawk strategies result in combat, although in theory, the escalated aggression will result in disproportionate injury because the animals will have unequivalent combative skills; hence, this is not truly a zero-sum game. Despite this, it is assumed that the value of the disputed resource is less than the cost of combat. Therefore, two Hawk strategies result in a minimax and two Dove strategies result in a maximin. In this situation, the pure strategy of Hawk-Dove will be preferred for each player, thereby resulting in two Nash equilibria for each conspecific (Hawk, Dove and Dove, Hawk). Table 2 reflects the outcome matrix for the Hawk-Dove strategies, where a 4 represents the greatest risk-reward payoff and a 1 reflects the lowest payoff.

Although game theorists are not necessarily interested in the outcome of the game as much as the strategy employed to solve for the maximum, it should be noted that pure strategies (i.e., always Dove or always Hawk) are not necessarily the most favorable approach to achieving optimization. In the case of Hawk-Dove, a mixed approach (i.e., randomization of the different strategies) is the most evolutionarily stable strategy in the long run.

Non–zero-sum games are frequently social dilemmas wherein private or individual interests are at odds with those of the collective. A classic two-person, non–zero-sum social dilemma is Prisoner's Dilemma. Prisoner's Dilemma is a strategic game in which the police concurrently interrogate two criminal suspects in separate rooms. In an attempt to collect more evidence supporting their case, the police strategically set each suspect in opposition. The tactic is to independently offer each prisoner the same deal. The Prisoner's Dilemma prisoner has two options: cooperate (i.e., remain silent) or defect (i.e., confess). Each prisoner's outcome is dependent not only on his or her behavior but the actions of his or her accomplice. If Prisoner A defects (confesses) while Prisoner B cooperates (remains silent), Prisoner A is freed and turns state's evidence, and Prisoner B receives a full 10-year prison sentence. In this scenario, the police have sufficient evidence to convict both prisoners on a lighter sentence without their shared confessions, so if Prisoners A and B both fail to confess, they both receive a 5-year sentence. However if both suspects confess, they each receive the full prison sentence of 10 years. Table 3 reflects the Prisoner's Dilemma outcome matrix, where the numeric values represent years in prison.

This example of Prisoner's Dilemma does contain a single Nash equilibrium (defect, defect), where both suspects optimize by betraying their accomplice, providing the police with a clear advantage in extracting confessions.

### *Non–Zero-Sum Cooperative Games*

Although game theory typically addresses situations in which players have conflicting interests, one way to maximize may be to modify one's strategy to compromise or cooperate to resolve the conflict. Non–zero-sum games within this category broadly include Tit-for-Tat, bargaining games, and coalition games. In non–zero-sum cooperative games, the emphasis is no longer on individual optimization; the maximum includes the optimization interests of other players or groups of players. This equalizes the distribution of resources among two or more players. In cooperative games, the focus shifts from more individualistic concept solutions to group solutions. In game theory parlance, players attempt to maximize their minimum loss, thereby selecting the

maximin and distributing the resources to one or more other players. Generally, these kinds of cooperative games occur between two or more groups that have a high probability of repeated interaction and social exchange. Broadly, non–zero-sum cooperative games use the principle of reciprocity to optimize the maximum for all players. If one player uses the maximin, the counter response should follow the principle of reciprocity and respond in kind. An example of this kind of non-zero-sum cooperative game is referred to as *Tit-for-Tat*.

Anatol Rapoport submitted Tit-for-Tat as a solution to a computer challenge posed by University of Michigan political science professor Robert Axelrod in 1980. Axelrod solicited the most renowned game theorists of academia to submit solutions for an iterated Prisoner's Dilemma, wherein the players (i.e., prisoners) were able to retaliate in response to the previous tactic of their opposing player (i.e., accomplice). Rapoport's Tit-for-Tat strategy succeeded in demonstrating optimization. Tit-for-Tat is a payback strategy typically between two players and founded on the principle of reciprocity. It begins with an initial cooperative action by the first player; henceforth, all subsequent actions reflect the last move of the second player. Thus, if the second player responds cooperatively (Tat), then the first player responds in kind (Tit) ad infinitum. Tit-for-Tat was not necessarily a new conflict resolution approach when submitted by Rapoport, but one favored historically as a militaristic strategy using a different name, equivalent retaliation, which reflected a Tit-for-Tat approach. Each approach is highly vulnerable; it is effective only as long as all players are infallible in their decisions. Tit-for-Tat fails as an optimum strategy in the event of an error. If a player makes a mistake and accidentally defects from the cooperative concept solution to a competitive action, then conflict ensues and the non–zero-sum cooperative game becomes one of competition. A variant of Tit-for-Tat, Tit-for-Two-Tats, is more effective in optimization as it reflects a magnanimous approach in the event of an accidental escalation. In this strategy, if one player errs through a competitive action, the second player responds with a cooperative counter, thereby inviting remediation to the first player. If, after the second Tat, the first player has not corrected his or her strategy back to one of cooperation, then the second player responds with a retaliatory counter.

Another type of non–zero-sum cooperative game falls within the class of games of negotiation. Although these are still games of conflict and strategy, as a point of disagreement exists, the game is the negotiation. Once the bargain is proffered and accepted, the game is over. The most simplistic of these games is the Ultimatum Game, wherein two players discuss the division of a resource. The first player proposes the *apportionment,* and the second player can either accept or reject the offer. The players have one turn at negotiation; consequently, if the first player values the resource, he or she must make an offer that is perceived, in theory, to be reasonable by the second player. If the offer is refused, there is no second trial of negotiations and neither player receives any of the resource. The proposal is actually an ultimatum of "take it or leave it." The Ultimatum Game is also a political game of power. The player who proposes the resource division may offer an unreasonable request, and if the second player maintains less authority or control, the lack of any resource may be so unacceptable that both players yield to the ultimatum. In contrast to the Ultimatum Game, bargaining games of alternating offers is a game of repeated negotiation and perfect information where all previous negotiations may be referenced and revised, and where the players enter a state of equilibrium through a series of trials consisting of offers and counteroffers until eventually an accord is reached and the game concludes.

A final example of a non–zero-sum cooperative game is a game of coalitions. This is a cooperative game between individuals within groups. Coalitional games are games of consensus, potentially through bargaining, wherein the player strategies are conceived and enacted by the coalitions. Both the individual players and their group coalitions have an interest in optimization. However, the existence of coalitions protects players from defecting individuals; the coalition maintains the power to initiate a concept solution and thus all plays of strategy.

*Heide Deditius Island*

*See also* Decision Rule; Probability, Laws of

### Further Readings

Aumann, R. J., & Brandenburger, A. (1995). Epistemic conditions for Nash equilibrium. *Econometrica, 63*(5), 1161–1180.
Axelrod, R. (1985). *The evolution of cooperation.* New York: Basic Books.
Barash, D. P. (2003). *The survival game: How game theory explains the biology of competition and cooperation.* New York: Owl Books.
Davis, M. D. (1983). *Game theory: A nontechnical introduction* (Rev. ed.). Mineola, NY: Dover.
Fisher, L. (2008). *Rock, paper, scissors: Game theory in everyday life*. New York: Basic Books.
Osborne, M. J., & Rubinstein, A. (1994). *A course in game theory.* Cambridge: MIT Press.
Savage, L. J. (1972). *The foundations of statistics* (2nd ed.). Mineola, NY: Dover.
Von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior.* Princeton, NJ: Princeton University Press.

# Gauss–Markov Theorem

The Gauss–Markov theorem specifies the conditions under which the ordinary least squares (OLS) estimator is also the best linear unbiased (BLU) estimator. Because these BLU estimator properties are guaranteed by the Gauss–Markov theorem under general conditions that are often encountered in practice, ordinary least squares has become what George Stigler describes as the "automobile of modern statistical analysis." Furthermore, many of the most important advances in regression analysis have been direct generalizations of ordinary least squares under the Gauss–Markov theorem to even more general conditions. For example, weighted least squares, generalized least squares, finite distributed lag models, first-differenced estimators, and fixed-effect panel models all extend the finite-sample results of the Gauss–Markov theorem to conditions beyond the classical linear regression model. After a brief discussion of the origins of the theorem, this entry further examines the Gauss–Markov theorem in the context of statistical estimation and regression analysis.

## Origins

In 1821, Carl Friedrich Gauss proved that the least squares method produced unbiased estimates that have the smallest variance, a result that has become the cornerstone of regression analysis. In his 1900 textbook on probability, Andrei Markov essentially rediscovered Gauss's theorem. By the 1930s, however, the result was commonly referred to as the Markov theorem rather than the Gauss theorem. Perhaps because of awareness of this misattribution to Markov, many statisticians today avoid using the Gauss or Markov label altogether, referring instead to the equivalence of ordinary least squares (OLS) estimator and best linear unbiased (BLU) estimator. Most econometricians, however, refer to the result instead by the compromise label used here, the Gauss–Markov theorem.

## Statistical Estimation and Regression

The goal of statistical estimation is to provide accurate guesses about parameters (statistical summaries) of a population from a subset or sample of the population. In regression estimation, the main population parameters of interest measure how changes in one (independent) variable influence the value of another (dependent) variable. Because statistical estimation involves estimating unknown population parameters from known sample data, there are actually two equations involved in any simple regression estimation: a population regression function, which is unknown and being estimated,

$$y = \alpha + \beta x + \mu$$

and a sample regression function, which serves as the estimator and is calculated from the available data in the sample,

$$y = \hat{\alpha} + \hat{\beta} x + \hat{\mu}$$

where

$y$ = dependent variable

$x$ = independent variable

$\alpha$ = $y$-intercept of population regression function

$\hat{\alpha}$ = $y$-intercept of sample regression function

$\beta$ = slope of population regression function

$\hat{\beta}$ = slope of sample regression function

$\mu$ = error, disturbance of population regression function

$\hat{\mu}$ = residual of sample regression function

Because statistical estimation involves the calculation of population parameters from a finite sample of data, there is always some uncertainty about how close statistical estimates are to actual population parameters. To sort out the many possible ways of estimating a population parameter from a sample of data, various properties have been proposed for estimators. The "ideal estimator" always takes the exact value of the population parameter it is estimating. This ideal is unachievable from a finite sample of the population, and estimation instead involves a trade-off between different forms of accuracy, such as unbiasedness and minimum variance. The best linear unbiased estimator, which is discussed next, represents such a trade-off.

## Best Linear Unbiased Estimation

Because the Gauss–Markov theorem uses the BLU estimator as its standard of optimality, the qualities of this estimator are described first.

### *Linear Parameters*

A parameter $\beta$ is defined as linear if it is a linear function of $y$, that is, $\beta = ay + b$, where $a$ and $b$ can be constants or functions of $x$, but not of $y$. Requiring that the parameters $\alpha$ and $\beta$ are linear does not preclude nonlinear relationships between $x$ and $y$, such as polynomial terms (e.g., $x^2$) and interaction terms (e.g., $x * z$) on the right-hand side of the regression equation and logarithmic terms on the left-hand side of the regression equation [e.g., $\ln(y)$]. This flexibility gives the Gauss–Markov theorem wide applicability, but there are also many important regression models that cannot be written in a form linear in the parameters. For example, the main estimators for dichotomous dependent variables, logit and probit, involve nonlinear parameters for which the Gauss–Markov theorem cannot be applied.

### *Unbiasedness $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$*

Because the estimates $\hat{\alpha}$ and $\hat{\beta}$ are calculated from a finite sample, they are likely to be different from the actual parameters $\alpha$ and $\beta$ calculated from the entire population. If the sample estimates are the same as the population parameters on average when a large number of random samples of the same size are taken, then the estimates are said to be unbiased.

### *Best (Minimum Variance)*

The estimates $\hat{\alpha}$ and $\hat{\beta}$ will also vary from sample to sample because each randomly drawn sample will contain different values of $x$ and $y$. In practice, typically, there is only one sample, and repeated random samples are drawn hypothetically to establish the sampling distribution of the sample estimates. A BLU estimator $\hat{\beta}$ is "best" in the sense that the sampling distribution of $\hat{\beta}$ has a smaller variance than any other linear unbiased estimator of $\beta$.

BLU estimation places a priority on unbiasedness, which must be satisfied before the minimum variance condition. However, if, instead of unbiasedness, we begin by requiring that the estimator meet some non−zero maximum threshold of variance, the estimator chosen may hinge in a critical way on this arbitrary choice of threshold. The BLU estimator is free of this kind of arbitrariness because it makes unbiasedness the starting point, and all estimators can be classified as either biased or unbiased. But there is a trade-off in choosing a BLU estimator, because variance may, in fact, be much more important than unbiasedness in some cases, and yet unbiasedness is always given priority over lower variance.

Maximum likelihood is an alternative estimation criterion that does not suffer from the limitation of subordinating lower variance to unbiasedness. Instead, the maximum likelihood estimator (MLE) is chosen by the simple and intuitive principle that the set of parameters should be that most likely to have generated the set of data in the sample. Some have suggested that this principle be used as the main starting point for most statistical estimation. However, the MLE has two distinct disadvantages to the BLU estimator. First, the calculation of the MLE requires that the distribution

of the errors be completely specified, whereas the Gauss–Markov theorem does not require full specification of the error distribution. Second, the maximum likelihood estimator offers only asymptotic (large sample) properties, whereas the properties of BLU estimators hold in finite (small) samples.

## Ordinary Least Squares Estimation

The ordinary least squares estimator calculates $\hat{\beta}$ by minimizing the sum of the squared residuals $\hat{\mu}$. However, without further assumptions, one cannot know how accurately OLS estimates $\beta$. These further assumptions are provided by the Gauss–Markov theorem.

The OLS estimator has several attractive qualities. First, the Gauss–Markov theorem ensures that it is the BLU estimator given that certain conditions hold, and these properties hold even in small sample sizes. The OLS estimator is easy to calculate and is guaranteed to exist if the Gauss–Markov asssumptions hold. The OLS regression line can also be intuitively understood as the expected value of $y$ for a given value of $x$. However, because OLS is calculated using squared residuals, it is also especially sensitive to outliers, which exert a disproportionate influence on the estimates.

## The Gauss–Markov Theorem

The Gauss–Markov theorem specifies conditions under which ordinary least squares estimators are also best linear unbiased estimators. Because these conditions can be specified in many ways, there are actually many different Gauss–Markov theorems. First, there is the theoretical ideal of necessary and sufficient conditions. These necessary and sufficient conditions are usually developed by mathematical statisticians and often specify conditions that are not intuitive or practical to apply in practice. For example, the most widely cited necessary and sufficient conditions for the Gauss–Markov theorem, which Simo Puntanen and George Styan refer to as "Zyskind's condition," states in matrix notation that a necessary and sufficient condition for OLS to be BLU with fixed $x$ and nonsingular variance-covariance (dispersion) matrix $\mathbf{\Omega}$ is the existence of a non-singular matrix $\mathbf{Q}$ satisfying the equation $\mathbf{\Omega}x = x\mathbf{Q}$. Because such complex necessary and sufficient conditions offer little intuition for assessing when a given model satisfies them, in applied work, looser sufficient conditions, which are easier to assess, are usually employed. Because these practical conditions are typically not also necessary conditions, it means they are generally stricter than is theoretically required for OLS to be BLU. In other words, there may be models that do not meet the sufficient conditions for OLS to be BLU, but where OLS is nonetheless BLU.

Now, this entry turns to the sets of sufficient conditions that are most commonly employed for two different types of regression models: (a) models where $x$ is fixed in repeated sampling, which is appropriate for experimental research, and (b) models where $x$ is allowed to vary from sample to sample, which is more appropriate for observational (nonexperimental) data.

### *Gauss–Markov Conditions for Experimental Research (Fixed* x*)*

In experimental studies, the researcher has control over the treatment administered to subjects. This means that in repeated experiments with the same size sample, the researcher would be able to ensure that the subjects in the treatment group get the same level of treatment. Because this level of treatment is essentially the value of the independent variable $x$ in a regression model, this is equivalent to saying that the researcher is able to hold $x$ fixed in repeated samples. This provides a much simpler data structure in experiments than is possible in observational data, where the researcher does not have complete control over the value of the independent variable $x$. The following conditions are sufficient to ensure that the OLS estimator is BLU when $x$ is fixed in repeated samples:

1. Model correctly specified
2. Regressors not perfectly collinear
3. $E(\mu) = 0$
4. Homoscedasticity
5. No serial correlation

### Model Correctly Specified

In practice, there are two main questions here: What variables should be in the model? What is the correct functional form of those variables? Omitting an important variable that is correlated with both the dependent variable and one or more independent variables necessarily produces biased estimates. On the other hand, including irrelevant variables, which may be correlated with the dependent variables or independent variables but not with both, does not bias OLS estimates. However, adding irrelevant variables is not without cost. It reduces the number of observations available for calculating the impact of each independent variable, and there have to be fewer variables than observations for least-squares estimates to exist. Specifying the incorrect functional form of an independent variable will also lead to biased regression estimates. For example, if a researcher leaves out a significant squared term of $x$, then the sample regression function will impose a linear relationship on what is actually a nonlinear relationship between $x$ and $y$.

### No Perfect Multicollinearity

Perfect multicollinearity occurs when two or more variables are simple linear functions of each other. Multiple regression calculates the effect of one variable while holding the other variables constant. But if perfect multicollinearity exists, then it is impossible to hold one variable constant with respect to the other variables with which it is perfectly correlated. Whenever one variable changes, the other variable changes by an exact linear function of the change in the first variable. Perfect multicollinearity does not typically cause problems in practice, because variables are rarely perfectly correlated unless they are simply different measures of the same construct. The problem of high but not perfect collinearity is more likely to be problematic in practice, although it does not affect the applicability of the Gauss–Markov theorem.

### $E(\mu) = 0$

The expected value of the errors must be zero in order for $\hat{\beta}$ to be an unbiased estimate of $\beta$. There is no way to directly test whether this assumption holds in the population regression function, but it will not typically be problematic whenever the model is correctly specified and a constant ($y$-intercept) term is included in the model specification.

### Homoscedastic Errors, $E(\mu^2) = \sigma^2$

The variance of the errors must be constant across observations. Consider a regression of the cost of automobiles purchased by consumers on the consumers' income. In general, one would expect a positive relationship so that people with higher incomes purchase more expensive automobiles on average. But one might also find that there is much more variability in the purchase price of cars for higher income consumers than for lower income consumers, simply because lower income consumers can afford a smaller range of vehicles. Thus, we expect heteroscedastic (nonconstant) errors in this case, and we cannot apply the Gauss–Markov theorem to OLS.

If the other Gauss–Markov assumptions apply, OLS still generates unbiased estimates of regression coefficients in the presence of heteroscedastic errors, but they are no longer the linear unbiased estimators with the minimum variance. Smaller variance estimates can be calculated by weighting observations according to the heteroscedastic errors, using an estimator called weighted least squares (WLS). When the researcher knows the exact form of the heteroscedasticity and the other Gauss–Markov conditions hold, the WLS estimator is BLU. The actual nature of the heteroscedasticity in the errors of the population regression function are usually unknown, but such weights can be estimated using the residuals from the sample regression function in a procedure known as feasible generalized least squares (FGLS). However, because FGLS requires an extra estimation step, estimating $\mu^2$ by $\hat{\mu}^2$, it no longer obtains BLU estimates, but estimates only with asymptotic (large sample) properties.

### No Serial Correlation, $E(\mu_i \mu_j) = 0$

The errors of different observations cannot be correlated with each other. Like the homoscedasticity assumption, the presence of serial correlation does not bias least-squares estimates but it does affect their efficiency. Serial correlation can be a more complex problem to treat than

heteroscedasticity because it can take many different forms. For example, in its simplest form, the error of one observation is correlated only with the error in the next observation. For such processes, Aitken's generalized least squares can be used to achieve BLU estimates if the other Gauss–Markov assumptions hold. If, however, errors are associated with the errors of more than one other observation at a time, then more sophisticated time-series models are more appropriate, and in these cases, the Gauss–Markov theorem cannot be applied. Fortunately, if the sample is drawn randomly, then the errors automatically will be uncorrelated with each other, so that there is no need to worry about serial correlation.

### *Gauss–Markov Assumptions for Observational Research (Arbitrary* x*)*

A parallel but stricter set of Gauss–Markov assumptions is typically applied in practice in the case of observational data, where the researcher cannot assume that $x$ is fixed in repeated samples.

1. Model correctly specified
2. Regressors not perfectly collinear
3. $E(\mu|x) = 0$
4. Homoscedastic errors, $E(\mu^2|x) = \sigma^2$
5. No serial correlation, $E(\mu_i\mu_j|x) = 0$

The first two assumptions are exactly the same as in the fixed-$x$ case. The other three sufficient conditions are augmented so that they hold conditional on the value of $x$.

#### $\mathbf{E}(\boldsymbol{\mu}|\mathbf{x}) = \mathbf{0}$

In contrast to the fixed-$x$ case, $E(\mu|x) = 0$ is a very strong assumption that means that in addition to having zero expectation, the errors are not associated with any linear or nonlinear function of $x$. In this case, $\hat{\beta}$ is not only unbiased, but also unbiased conditional on the value of $x$, a stronger form of unbiasedness than is strictly needed for BLU estimation. Indeed, there is some controversy about whether this assumption must be so much stronger than in the fixed-$x$ case, given that the model is correctly specified. Nevertheless, $E(\mu|x) = 0$ is typically used in applying the Gauss–Markov theorem to observational data, perhaps to reiterate the potential specification problem of omitted confounding variables when there is not random assignment to treatment and control groups.

#### Homoscedastic Errors, $E(\boldsymbol{\mu}^2|\boldsymbol{x}) = \boldsymbol{\sigma}^2$

The restriction on heteroscedasticity of the errors is also strengthened in the case where $x$ is not fixed. In the fixed-$x$ case, the error of each observation was required to have the same variance. In the arbitrary-$x$ case, the errors are also required to have the same variance across all possible values of $x$. This is tantamount to requiring that the variance of the errors not be a (linear or nonlinear) function of $x$. Again, as in the fixed-$x$ case, violations of this assumption will still yield unbiased estimates of regression coefficients as long as the first three assumptions hold. But such heteroscedasticity will yield inefficient estimates unless the heteroscedasticity is addressed in the way discussed in the fixed-$x$ section.

#### No Serial Correlation, $E(\boldsymbol{\mu}_i\boldsymbol{\mu}_j|\boldsymbol{x}) = 0$

Finally, the restriction on serial correlation in the errors is strengthened to prohibit serial correlation that may be a (linear or nonlinear) function of $x$. Violations of this assumption will still yield unbiased least-squares estimates, but these estimates will not have the minimum variance among all unbiased linear estimators. In particular, it is possible to reduce the variance by taking into account the serial correlation in the weighting of observations in least squares. Again, if the sample is randomly drawn, then the errors will automatically be uncorrelated with each other. In time-series data in particular, it is usually inappropriate to assume that the data are drawn as a random sample, so special care must be taken to ensure that $E(\mu_i\mu_j|x) = 0$ before employing least squares. In most cases, it will be inapproriate to use least squares for time-series data. However, F. W. McElroy has provided a useful set of necessary and sufficient conditions for the Gauss–Markov theorem. For models with a $y$-intercept and a very simple form of serial correlation known as exchangeability, the OLS estimator is still the best linear unbiased estimator. A useful necessary and sufficient condition for the Gauss–Markov theorem in the

case of arbitrary $x$ is provided by McElroy. If the model includes an intercept term, McElroy shows that OLS will still be BLU in the presence of a weak form of serial correlation where all of the errors are equally correlated with each other.

*Roger Larocca*

***See also*** Biased Estimator; Estimation; Experimental Design; Homoscedasticity; Least Squares, Methods of; Observational Research; Regression Coefficient; Regression to the Mean; Residuals; Serial Correlation; Standard Error of Estimate; Unbiased Estimator

### Further Readings

Berry, W. D. (1993). *Understanding regression assumptions*. Newbury Park, CA: Sage.

McElroy, F. W. (1967). A necessary and sufficient condition that ordinary least-squares estimators be best linear unbiased estimators. *Journal of the American Statistical Association, 62,* 1302–1304.

Plackett, R. L. (1949). A historical note on the method of least squares. *Biometrika, 36,* 458–460.

Puntanen, S., & Styan, G. P. H. (1989). The equality of the ordinary least squares estimator and the best linear unbiased estimator. *American Statistician,* 43, 153–161.

Stigler, S. M. (1980). Gauss and the invention of least squares. *Annals of Statistics, 9,* 465–474.

## Generalizability Theory

Generalizability theory (G theory), originally developed by Lee J. Cronbach and his associates, is a measurement theory that provides both a conceptual framework and a set of statistical procedures for a comprehensive analysis of test reliability. Building on and extending classical test theory (CTT) and analysis of variance (ANOVA), G theory provides a flexible approach to modeling measurement error for different measurement conditions and types of decisions made based on test results. This entry introduces the reader to the basics of G theory, starting with the advantages of G theory, followed by key concepts and terms and some illustrative examples representing different G-theory analysis designs.

### Advantages

There are a few approaches to the investigation of test reliability, that is, the consistency of measurement obtained in testing. For example, for norm-referenced testing (NRT), CTT reliability indexes show the extent to which candidates are rank-ordered consistently across test tasks, test forms, occasions, and so on (e.g., Cronbach's alpha and parallel-form and test-retest reliability estimates). In contrast, in criterion-referenced testing (CRT), various statistics are used to examine the extent to which candidates are consistently classified into different categories (score or ability levels) across test forms, occasions, test tasks, and so on. Threshold-loss agreement indexes such as the agreement coefficient and the kappa coefficient are some examples.

Why might one turn to G theory despite the availability of these different approaches to reliability investigation? G theory is a broadly defined analytic framework that addresses some limitations of the traditional approaches. First, the approaches above address only NRT *or* CRT, whereas G theory accommodates both (called *relative decisions* and *absolute decisions,* respectively), yielding measurement error and reliability estimates tailored to the specific type of decision making under consideration. Second, CTT reliability estimates take account of only one source of measurement error at a time. Thus, for example, when one is concerned about the consistency of examinee rank-ordering across two testing occasions and across different raters, he or she needs to calculate two separate CTT reliability indexes (i.e., test-retest and interrater reliability estimates). In contrast, G theory provides reliability estimates accounting for both sources of error simultaneously. The G theory capability to analyze multiple sources of error within a single analysis is particularly useful for optimizing the measurement design to achieve an acceptable level of measurement reliability.

### Key Concepts and Terms

A fundamental concept in G theory is *dependability.* Dependability is defined as the extent to which the generalization one makes about a given candidate's *universe score* based on an observed test

score is accurate. The universe score is a G-theory analogue of the true score in CTT and is defined as the average score a candidate would have obtained across an infinite number of testing under measurement conditions that the investigator is willing to accept as exchangeable with one another (called *randomly parallel* measures). Suppose, for example, that an investigator has a large number of vocabulary test items. The investigator might feel comfortable treating these items as randomly parallel measures because trained item writers have carefully developed these items to target a specific content domain, following test specifications. The employment of randomly parallel measures is a key assumption of G theory. Note the difference of this assumption from the CTT assumption, where sets of scores that are involved in a reliability calculation must be statistically parallel measures (i.e., two sets of scores must share the same mean, the same standard deviation, and the same correlation to a third measure).

Observed test scores can vary for a number of reasons. One reason may be the true differences across candidates in terms of the ability of interest (called the *object of measurement*). Other reasons may be the effects of different sources of measurement error: Some are systematic (e.g., item difficulty), whereas others are unsystematic (e.g., fatigue). In estimating a candidate's universe score, one cannot test the person an infinite number of times in reality. Therefore, one always has to estimate a candidate's universe score based on a limited number of measurements available. In G theory, a systematic source of variability that may affect the accuracy of the generalization one makes is called a *facet*. There are two types of facets. A facet is *random* if the intention is to generalize beyond the conditions actually used in an assessment. In this case, measurement conditions are conceptualized as a representative sample of a much larger population of admissible observations (called the *universe* in G theory). Alternatively, a facet is *fixed* when there is no intention to generalize beyond the conditions actually used in the assessment, because either the set of measurement conditions exhausts all admissible observations in the universe or the investigator has chosen specific conditions on purpose.

Different sources of measurement error are analyzed in a two-step procedure. The first step is a generalizability study (G study), where the observed score variance is decomposed into pieces attributable to different sources of score variability called *variance components* associated with a facet(s) identified by the investigator.

As shown in detail in the numerical example below, G-study variance component estimates are typically obtained by fitting a random-effects ANOVA model to data. The primary purpose of this analysis is to obtain mean squares for different effects that are needed for the calculation of variance component estimates. Variance component estimates are key building blocks of G theory. A G-study variance component estimate indicates the magnitude of the effect of a given source of variability on the observed score variance for a hypothetical measurement design where only a single observation is used for testing (e.g., a test consisting of one item).

The G-study variance component estimates are then used as the baseline data in the second step of the analysis called a decision study (D study). In a D study, variance components and measurement reliability can be estimated for a variety of hypothetical measurement designs (for instance, a test consisting of multiple items) and types of score interpretations of interest.

## A Numerical Example: One-Facet Crossed Study Design

Suppose that an investigator wants to analyze results of a grammar test consisting of 40 items administered to 60 students in a French language course. Because these items have been randomly selected from a large pool of items, the investigator defines items as a random facet. In this test, all candidates (persons) complete all items. In G-theory terms, this study design is called a one-facet study design because it involves only one facet (items). Moreover, persons and items are called *crossed* because for each person, scores for all items are available (denoted $p \times i$, where the " $\times$ " is read "crossed with").

For this one-facet crossed study design, the observed score variance is decomposed into three variance components:

1. Person variance component [$\sigma^2(p)$]: The observed score variance due to the true

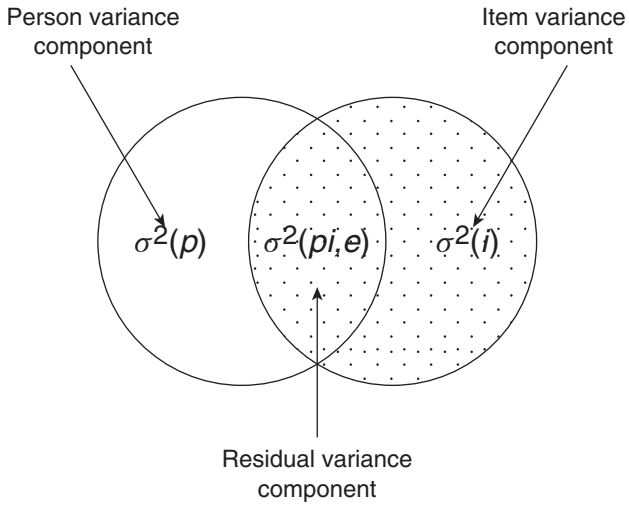



**Figure 1** Decomposition of the Observed Score Variance Into Variance Components in a One-Facet Crossed Study Design

differences among candidates on the target ability

2. Item variance component [$\sigma^2(i)$]: The observed score variance due to the differences across items in terms of difficulty
3. Residual variance component [$\sigma^2(pi,e)$]: The observed score variance due to a combination of two confounded sources of error: (a) the observed score variance due to an interaction between persons and items, that is, the extent to which the rank-ordering of persons differs from one item to another, and (b) the score variance due to undifferentiated error, consisting of other systematic sources of variance (facets) not taken account of in this study design and random sources of error (e.g., fatigue).

Figure 1 schematically represents the decomposition of the observed score variance into the person ($p$), item ($i$), and residual ($pi,e$) variance components. Note that the person variance component is a source of variability due to the object of measurement. The item and residual variance components (dotted areas in the figure) reflect sources of variability associated with the facet of measurement.

Calculating variance components requires a set of formulas called expected mean square (EMS) equations for a specific study design. Below are the equations for the one-facet crossed study design:

**Table 1** ANOVA Table for a One-Facet Crossed Study Design

| *Source* | *Degrees of Freedom (df)* | *Sum of Squares (SS)* | *Mean Squares (MS)* |
|---|---|---|---|
| Persons ($p$) | 59 | 96.943 | 1.643 |
| Items ($i$) | 39 | 173.429 | 4.457 |
| *pi,e* | 2,301 | 416.251 | 0.181 |
| Total | 2,399 | 686.623 | |

$$EMS(p) = \sigma^2(pi,e) + n_i\sigma^2(p) \quad (1)$$

$$EMS(i) = \sigma^2(pi,e) + n_p\sigma^2(i) \quad (2)$$

$$EMS(pi,e) = \sigma^2(pi,e) \quad (3)$$

Here, $n$ denotes the G-study sample size. Theoretically, an EMS is defined as the average mean square value across repeated analyses of samples from the same population of examinees and the same universes of admissible observations for the same study design. Because an EMS is usually unknown, the EMSs in these equations are replaced with the observed mean square values for persons ($p$), items ($i$), and residuals ($pi,e$) obtained from an ANOVA of sample data. Thus, it should be remembered that the obtained variance components are estimates obtained from a sample. As an example, Table 1 shows what would be obtained from a random-effects ANOVA of the data for this example based on the 60 students' responses to the 40 grammar items.

The variance component estimates are obtained by solving the equations for $\sigma^2(p)$, $\sigma^2(i)$, and $\sigma^2(pi,e)$, respectively:

$$1.643 = 0.181 + 40 \times \sigma^2(p) \quad \sigma^2(p) = 0.037$$
$$4.447 = 0.181 + 60 \times \sigma^2(i) \quad \sigma^2(i) = 0.071$$
$$0.181 = \sigma^2(pi,e) \quad \sigma^2(pi,e) = 0.181$$

Then, the obtained G-study variance component estimates can be analyzed further by preparing a table like Table 2.

The left panel of Table 2 provides the magnitudes of the G-study variance component estimates for persons, items, and residuals, along with the percentage of the observed score variance explained by each source of score variability for

**Table 2** Variance Component Estimates for a One-Facet Crossed Study Design

| | *G-Study Variance Components (for a single observation)* | | *D-Study Variance Components (for 50 items)* |
|---|---|---|---|
| *Source* | *Estimated Variance Components* | *Percentage of Total Variance* | *Estimated Variance Components* |
| Persons ($p$) | 0.037 | 12.8 | 0.037 |
| Items ($i$) | 0.071 | 24.6 | $\sigma^2(i)/n_i' = 0.071/50 = 0.001$ |
| $pi,e$ | 0.181 | 62.6 | $\sigma^2(pi,e)/n_i' = 0.181/50 = 0.004$ |
| Total | 0.289 | 100.0 | |

a single observation. As can be seen in the table, the person, item, and residual variance components account for 12.8%, 24.6%, and 62.6% of the total score variance, respectively.

Based on the G-study results above, a D study can be conducted to estimate score reliability for an alternative measurement design. As in CTT, where the Spearman-Brown prophecy formula is used to estimate test reliability for different test lengths, one can estimate the measurement reliability for a test involving different numbers of items. As an example, the right panel of Table 2 shows the D-study results for 50 items. First, D-study variance component estimates for this measurement design are obtained by dividing the G-study variance component estimates associated with the facet of measurement [i.e., $\sigma^2(i)$ and $\sigma^2(pi,e)$ in this case] by the D-study sample size for the item facet ($n_i' = 50$).

Second, a summary index of reliability, similar to what one might obtain in a CTT analysis, can be calculated for the 50-item scenario. G theory provides two types of reliability-like indexes for different score interpretations: a *generalizability coefficient* (denoted $E\rho^2$) for relative decisions, and an *index of dependability* (denoted $\phi$; often called phi coefficient) for absolute decisions. These coefficients are obtained in two steps. First, the error variance appropriate for the type of decision is calculated. For relative decisions, all variance components involving persons, except the object of measurement [$\sigma^2(p)$], contribute to the *relative error variance* [$\sigma^2(Rel)$]. Thus, for this one-facet crossed study example, only the residual variance component [$\sigma^2(pi,e)$] contributes to the relative error variance; hence, $\sigma^2(Rel) = \sigma^2(pi,e) = 0.004$. For absolute decisions, all variance components except that for the object of measurement [$\sigma^2(p)$] contribute to the *absolute error variance* [$\sigma^2(Abs)$]. In this example, both $\sigma^2(i)$ and $\sigma^2(pi,e)$ will contribute to the absolute error variance. Thus, $\sigma^2(Abs) = \sigma^2(i) + \sigma^2(pi,e) = 0.001 + 0.004 = 0.005$. Second, a G-coefficient or a phi-coefficient is obtained by dividing the variance component due to the object of measurement [$\sigma^2(p)$], which is also called the *universe-score variance,* by the sum of itself and the appropriate type of error variance. Thus, for this one-facet crossed study example, the G- and phi-coefficients are calculated as follows:

$$\begin{aligned} E\rho^2 &= \sigma^2(p)/[\sigma^2(p) + \sigma^2(Rel)] \\ &= 0.037/(0.037 + 0.004) = 0.902 \\ \phi &= \sigma^2(p)/[\sigma^2(p) + \sigma^2(Abs)] \\ &= 0.037/(0.037 + 0.005) = 0.881 \end{aligned}$$

G theory is conceptually related to CTT. Under certain conditions, CTT and G theory analysis results yield identical results. This is the case when a one-facet crossed study design is employed for relative decisions. Thus, for example, the G-coefficient obtained from the one-facet D study with 50 items above is identical to Cronbach's alpha for the same number of items.

## Other Study Designs

The numerical example above is one of the simplest designs that can be implemented in a G-theory data analysis. Below are some examples of crossed study designs involving multiple random facets as well as other study designs involving a nested facet or a fixed facet.

### *A Crossed Study Design With Two Random Facets*

As mentioned above, one can take advantage of the strength of G theory when multiple sources of error are modeled simultaneously. Suppose, for example, in a speaking test each student completes three items, and two raters score each student's responses to all three items. In this case, the investigator may identify two facets: items and raters. Persons, items, and raters are crossed with one another because (a) all students complete all items, (b) all students are rated by both raters, and (c) both raters score student responses to all items. This study design is called a two-facet crossed study design ($p \times r \times i$).

### *A Study Design Involving a Nested Facet*

Some G theory analyses may be conducted for study designs involving a nested facet. Facet A is nested within Facet B if different, multiple levels of Facet A are associated with each level of Facet B (Shavelson & Webb). Typically, a nested facet is found in two types of situations. The first is when one facet is nested within another facet by definition. A common example is a reading test consisting of groups comprehension items based on different passages. In this case, the item facet is nested within the reading passage facet because a specific group of items is associated only with a particular passage. The second is a situation where one chooses to use a nested study design, although employing a crossed study design is possible. For instance, collecting data for the two-facet crossed study design example above can be resource intensive because all raters have to score all student responses. In this case, a decision might be made to have different rater pairs score different items to shorten the scoring time. This results in a two-facet study design where raters are nested within items [denoted $p \times (r : i)$, where the ":" is read "nested within"].

### *Study Designs Involving a Fixed Facet*

Because G theory is essentially a measurement theory for modeling random effects, at least one facet identified in a study design must be a random effect. Multifacet study designs may involve one or more fixed facets, however. Suppose, in the speaking test described earlier in the two-facet crossed study example, that each candidate response is evaluated on three dimensions: pronunciation, grammar, and fluency. These dimensions can be best conceptualized as the levels in a fixed facet because they have been selected as the scoring criteria on purpose.

There are some alternatives to model such fixed facets in G theory. Whichever approach is employed, the decision for selecting an approach must be made based on careful considerations of various substantive issues. One approach is to conduct a two-facet crossed study ($p \times r \times i$) for each dimension separately. This approach is preferred if the investigator believes that the three dimensions are conceptually so different that study results cannot be interpreted meaningfully at the aggregated level, or if the variance component estimates vary widely across the dimensions. Alternatively, one can analyze all dimensions simultaneously by conducting a three-facet crossed study ($p \times r \times i \times d$, where dimensions, or $d$, are treated as a fixed facet). This approach is reasonable if variance component estimates averaged across the dimensions can offer meaningful information for a particular assessment context, or if the variance component estimates obtained from separate analyses of the dimensions in the $p \times r \times i$ study runs are similar across the dimensions.

Another possible approach is to use multivariate G theory. Although multivariate G theory is beyond the scope of this introduction to G theory, Robert Brennan's 2001 volume in the Further Readings list provide an extensive discussion of this topic.

## Computer Programs

Computer programs specifically designed for G theory analyses offer comprehensive output for both G and D studies for a variety of study designs. Brennan's GENOVA Suite offers three programs: GENOVA, urGENOVA, and mGENOVA. GENOVA and urGENOVA handle different study designs for univariate G-theory analyses, on which this entry has focused, whereas mGENOVA is designed for multivariate G-theory analyses.

*Yasuyo Sawaki*

*See also* Analysis of Variance (ANOVA); Classical Test Theory; Coefficient Alpha; Interrater Reliability; Random Effects Models; Reliability

### Further Readings

Brennan, R. L. (1992). *Elements of generalizability theory.* Iowa City, IA: ACT.
Brennan, R. L. (2001). *Generalizability theory.* New York: Springer-Verlag.
Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.
Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage.
Webb, N. M., & Shavelson, R. J. (1981). Multivariate generalizability of general educational development ratings. *Journal of Educational Measurement, 18*(1), 13–22.
Webb, N. M., Shavelson, R. J., & Maddahian, E. (1983). Multivariate generalizability theory. In L. J. Fyans, Jr. (Ed.), *Generalizability theory: Inferences and practical applications* (pp. 67–81). San Francisco: Jossey-Bass.

## General Linear Model

The general linear model (GLM) provides a general framework for a large set of models whose common goal is to explain or predict a quantitative dependent variable by a set of independent variables that can be categorical or quantitative. The GLM encompasses techniques such as Student's *t* test, simple and multiple linear regression, analysis of variance, and covariance analysis. The GLM is adequate only for *fixed*-effect models. In order to take into account *random*-effect models, the GLM needs to be extended and becomes the *mixed*-effect model.

### Notations

Vectors are denoted with boldface lower-case letters (e.g., $\mathbf{y}$), and matrices are denoted with boldface upper-case letters (e.g., $\mathbf{X}$). The transpose of a matrix is denoted by the superscript $^{\mathrm{T}}$, and the inverse of a matrix is denoted by the superscript $^{-1}$. There are $I$ observations. The values of a quantitative dependent variable describing the $I$ observations are stored in an $I$ by 1 vector denoted $\mathbf{y}$. The values of the independent variables describing the $I$ observations are stored in an $I$ by $K$ matrix denoted $\mathbf{X}$. $K$ is smaller than $I$, and $\mathbf{X}$ is assumed to have rank $K$ (i.e., $\mathbf{X}$ is full rank on its columns). A quantitative independent variable can be directly stored in $\mathbf{X}$, but a qualitative independent variable needs to be recoded with as many columns as there are degrees of freedom for this variable. Common coding schemes include dummy coding, effect coding, and contrast coding.

### *Core Equation*

For the GLM, the values of the dependent variable are obtained as a *linear* combination of the values of the independent variables. The vectors for the coefficients of the linear combination are stored in a $K$ by 1 vector denoted $\mathbf{b}$. In general, the values of $\mathbf{y}$ cannot be perfectly obtained by a linear combination of the columns of $\mathbf{X}$, and the difference between the actual and the predicted values is called the *prediction error*. The values of the error are stored in an $I$ by 1 vector denoted $\mathbf{e}$. Formally, the GLM is stated as

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e}. \tag{1}$$

The predicted values are stored in an $I$ by 1 vector denoted $\hat{\mathbf{y}}$, and therefore, Equation 1 can be rewritten as

$$\mathbf{y} = \hat{\mathbf{y}} + \mathbf{e} \quad \textit{with} \quad \hat{\mathbf{y}} = \mathbf{Xb}. \tag{2}$$

Putting together Equations 1 and 2 shows that

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}. \tag{3}$$

### *Additional Assumptions*

The independent variables are assumed to be fixed variables (i.e., their values will not change for a replication of the experiment analyzed by the GLM, and they are measured without error). The error is interpreted as a *random* variable; in addition, the $I$ components of the error are assumed to be independently and identically distributed (i.i.d.), and their distribution is assumed to be a normal distribution with a zero mean and a variance denoted $\sigma^2_{\mathbf{e}}$. The values of the dependent variable are assumed to be a random sample of a population of interest. Within this framework,

the vector $\mathbf{b}$ is seen as an estimation of the population parameter vector $\boldsymbol{\beta}$.

## Least Square Estimate

Under the assumptions of the GLM, the population parameter vector $\boldsymbol{\beta}$ is estimated by $\mathbf{b}$, which is computed as

$$\mathbf{b} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}. \tag{4}$$

This value of $\mathbf{b}$ minimizes the residual sum of squares (i.e., $\mathbf{b}$ is such that $\mathbf{e}^{\mathrm{T}}\mathbf{e}$ is minimum).

### *Sums of Squares*

The total sum of squares of $\mathbf{y}$ is denoted $SS_{\text{total}}$, and it is computed as

$$SS_{\text{total}} = \mathbf{y}^{\mathrm{T}}\mathbf{y}. \tag{5}$$

Using Equation 2, the total sum of squares can be rewritten as

$$\begin{aligned} SS_{\text{total}} &= \mathbf{y}^{\mathrm{T}}\mathbf{y} = (\hat{\mathbf{y}} + \mathbf{e})^{\mathrm{T}}(\hat{\mathbf{y}} + \mathbf{e}) \\ &= \hat{\mathbf{y}}^{\mathrm{T}}\hat{\mathbf{y}} + \mathbf{e}^{\mathrm{T}}\mathbf{e} + 2\hat{\mathbf{y}}^{\mathrm{T}}\mathbf{e}, \end{aligned} \tag{6}$$

but it can be shown that $2\hat{\mathbf{y}}^{\mathrm{T}}\mathbf{e} = 0$, and therefore, Equation 6 becomes

$$SS_{\text{total}} = \mathbf{y}^{\mathrm{T}}\mathbf{y} = \hat{\mathbf{y}}^{\mathrm{T}}\hat{\mathbf{y}} + \mathbf{e}^{\mathrm{T}}\mathbf{e}. \tag{7}$$

The first term of Equation 7 is called the *model* sum of squares and is denoted $SS_{\text{Model}}$. It is equal to

$$SS_{\text{model}} = \hat{\mathbf{y}}^{\mathrm{T}}\hat{\mathbf{y}} = \mathbf{b}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{b}. \tag{8}$$

The second term of Equation 7 is called the *residual* or the *error* sum of squares and is denoted $SS_{\text{residual}}$. It is equal to

$$SS_{\text{residual}} = \mathbf{e}^{\mathrm{T}}\mathbf{e} = (\mathbf{y} - \mathbf{X}\mathbf{b})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\mathbf{b}). \tag{9}$$

### *Sampling Distributions of the Sums of Squares*

Under the assumptions of normality and i.i.d. for the error, the ratio of the residual sum of squares to the error variance $\frac{SS_{\text{residual}}}{\sigma_{\mathrm{e}}^2}$ is distributed as a $\chi^2$ with a number of degrees of freedom of $\nu = I - K - 1$. This is abbreviated as

$$\frac{SS_{\text{residual}}}{\sigma_{\mathrm{e}}^2} \sim \chi^2(\nu). \tag{10}$$

By contrast, the ratio of the model sum of squares to the error variance $\frac{SS_{\text{model}}}{\sigma_{\mathrm{e}}^2}$ is distributed as a *noncentral* $\chi^2$ with $\nu = K$ degrees of freedom and noncentrality parameter

$$\lambda = \frac{2}{\sigma_{\mathrm{e}}^2}\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}.$$

This is abbreviated as

$$\frac{SS_{\text{model}}}{\sigma_{\mathrm{e}}^2} \sim \chi^2(\nu,\ \lambda). \tag{11}$$

From Equations 10 and 11, it follows that the ratio

$$\begin{aligned} F &= \frac{SS_{\text{model}}/\sigma_{\mathrm{e}}^2}{SS_{\text{residual}}/\sigma_{\mathrm{e}}^2} \times \frac{I - K - 1}{K} \\ &= \frac{SS_{\text{model}}}{SS_{\text{residual}}} \times \frac{I - K - 1}{K} \end{aligned} \tag{12}$$

is distributed as a noncentral Fisher's $F$ with $\nu_1 = K$ and $\nu_2 = I - K - 1$ degrees of freedom and noncentrality parameter equal to

$$\lambda = \frac{2}{\sigma_{\mathrm{e}}^2}\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}.$$

In the specific case when the null hypothesis of interest states that $H_0 : \boldsymbol{\beta} = \mathbf{0}$, the noncentrality parameter vanishes and then the $F$ ratio from Equation 12 follows a standard (i.e., central) Fisher's distribution with $\nu_1 = K$ and $\nu_2 = I - K - 1$ degrees of freedom.

## Test on Subsets of the Parameters

Often, one is interested in testing only a subset of the parameters. When this is the case, the $I$ by $K$ matrix $\mathbf{X}$ can be interpreted as composed of two blocks: an $I$ by $K_1$ matrix $\mathbf{X}_1$ and an $I$ by $K_2$ matrix $\mathbf{X}_2$ with $K = K_1 + K_2$. This is expressed as

$$\mathbf{X} = [\mathbf{X}_1 \vdots \mathbf{X}_2]. \tag{13}$$

Vector $\mathbf{b}$ is partitioned in a similar manner as

$$\mathbf{b} = \begin{bmatrix} \mathbf{b}_1 \\ \cdots \\ \mathbf{b}_2 \end{bmatrix}. \tag{14}$$

In this case, the model corresponding to Equation 1 is expressed as

$$\begin{aligned} \mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e} &= [\mathbf{X}_1 \vdots \mathbf{X}_2] \begin{bmatrix} \mathbf{b}_1 \\ \cdots \\ \mathbf{b}_2 \end{bmatrix} + \mathbf{e} \\ &= \mathbf{X}_1\mathbf{b}_1 + \mathbf{X}_2\mathbf{b}_2 + \mathbf{e} \end{aligned} \tag{15}$$

For convenience, we will assume that the test of interest concerns the parameters $\boldsymbol{\beta}_2$ estimated by vector $\mathbf{b}_2$ and that the null hypothesis to be tested corresponds to a *semipartial* hypothesis, namely, that adding $\mathbf{X}_2$ *after* $\mathbf{X}_1$ does not improve the prediction of $\mathbf{y}$. The first step is to evaluate the quality of the prediction obtained when using $\mathbf{X}_1$ alone. The estimated value of the parameters is denoted $\widetilde{\mathbf{b}}_1$—a new notation is needed because in general, $\mathbf{b}_1$ is different from $\widetilde{\mathbf{b}}_1$ ($\mathbf{b}_1$ and $\widetilde{\mathbf{b}}_1$ are equal only if $\mathbf{X}_1$ and $\mathbf{X}_2$ are two orthogonal blocks of columns). The model relating $\mathbf{y}$ to $\mathbf{X}_1$ is called a *reduced* model. Formally, this reduced model is obtained as

$$\mathbf{y} = \mathbf{X}_1\widetilde{\mathbf{b}}_1 + \widetilde{\mathbf{e}}_1 \tag{16}$$

(where $\widetilde{\mathbf{e}}_1$ is the error of prediction for the reduced model). The model sum of squares for the reduced model is denoted $SS_{\widetilde{\mathbf{b}}_1}$ [see Equation (9) for its computation]. The semipartial sum of squares for $\mathbf{X}_2$ is the sum of squares *over and above* the sum of squares already explained by $\mathbf{X}_1$. It is denoted $SS_{\mathbf{b}_2|\mathbf{b}_1}$ and it is computed as

$$SS_{\mathbf{b}_2|\mathbf{b}_1} = SS_{\text{model}} - SS_{\widetilde{\mathbf{b}}_1} \tag{17}$$

The null hypothesis test indicating that $\mathbf{X}_2$ does not improve the prediction of $\mathbf{y}$ over and above $\mathbf{X}_1$ is equivalent to testing the null hypothesis that $\mathbf{b}_2$ is equal to 0. It can be tested by computing the following $F$ ratio:

$$F_{\mathbf{b}_2|\mathbf{b}_1} = \frac{SS_{\mathbf{b}_2|\mathbf{b}_1}}{SS_{\text{residual}}} \times \frac{I - K - 1}{K_2}. \tag{18}$$

When the null hypothesis is true, $F_{\mathbf{b}_2|\mathbf{b}_1}$ follows a Fisher's $F$ distribution with $\nu_1 = K_2$ and $\nu_2 = I - K - 1$ degrees of freedom, and therefore, $F_{\mathbf{b}_2|\mathbf{b}_1}$ can be used to test the null hypothesis that $\boldsymbol{\beta}_2 = \mathbf{0}$.

## Specific Cases

The GLM comprises several standard statistical techniques. Specifically, linear regression is obtained by augmenting the matrix of independent variables by a column of ones (this additional column codes for the intercept). Analysis of variance is obtained by coding the experimental effect in an appropriate way. Various schemes can be used, such as effect coding, dummy coding, or contrast coding (with as many columns as there are degrees of freedom for the source of variation considered). Analysis of covariance is obtained by combining the quantitative independent variables expressed as such and the categorical variables expressed in the same way as for an analysis of variance.

## Limitations and Extensions

The general model, despite its name, is not completely general and has several limits that have spurred the development of "generalizations" of the general linear model. Some of the most notable limits and some palliatives are listed below.

The general linear model requires $\mathbf{X}$ to be full rank, but this condition can be relaxed by using (cf. Equation 4) the Moore-Penrose generalized inverse (often denoted $\mathbf{X}^+$ and sometimes called a "pseudo-inverse") in lieu of $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. Doing so, however, makes the problem of estimating the model parameters more delicate and requires the use of the notion of *estimable* functions.

The general linear model is a *fixed*-effects model, and therefore, it does not naturally work with random-effects models (including multifactorial repeated or partially repeated measurement designs). In this case (at least for balanced designs), the sums of squares are computed correctly, but the $F$ tests are likely to be incorrect. A palliative to this problem is to compute expected values for the different sums of squares and to compute $F$ tests accordingly. Another, more general, approach is to model separately the fixed

effects and the random effects. This is done with mixed-effects models.

Another obvious limit of the general *linear* model is to model only linear relationships. In order to include some nonlinear models (such as logistic regression), the GLM needs to be extended to the class of the *generalized* linear models.

*Hervé Abdi*

***See also*** Analysis of Variance (ANOVA); Analysis of Covariance (ANCOVA); Contrast Analysis; Degrees of Freedom; Dummy Coding; Effect Coding; Experimental Design; Fixed-Effects Models; Gauss–Markov Theorem; Homoscedasticity; Least Squares, Methods of; Matrix Algebra; Mixed Model Design; Multiple Regression; Normality Assumption; Random Error; Sampling Distributions; Student's *t* Test; *t* Test, Independent Samples; *t* Test, One Sample; *t* Test, Paired Samples

**Further Readings**

Abdi, H., Edelman, B., Valentin, D., & Dowling, W. J. (2009). *Experimental design and analysis for psychology*. Oxford, UK: Oxford University Press.

Brown, H., & Prescott, R. (2006). *Applied mixed models in medicine*. London: Wiley.

Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin,* 70, 426–443.

Fox, J. (2008). *Applied regression analysis and generalized linear models*. Thousand Oaks, CA: Sage.

Graybill, F. A. (1976). *Theory and application of the linear model*. North Scituate, MA: Duxbury.

Hocking, R. R. (2003). *Methods and applications of linear models*. New York: Wiley.

Rencher, A. C., & Schaafe, G. B. (2008). *Linear models in statistics*. New York: Wiley.

Searle, S. R. (1971). *Linear models*. New York: Wiley.

# Graphical Display of Data

Graphs and charts are now a fundamental component of modern research and reporting. Today, researchers use many graphical means such as histograms, box plots, and scatterplots to better understand their data. Graphs are effective for displaying and summarizing large amounts of numerical data, and are useful for showing trends, patterns, and relationships between variables. Graphs and charts are also used to enhance reporting and communication. Graphical displays often provide vivid color and bring life to documents, while also simplifying complex narrative and data. This entry discusses the importance of graphs, describes common techniques for presenting data graphically, and provides information on creating effective graphical displays.

## Importance of Graphs in Data Analysis and Reporting

In the conduct of research, researchers accumulate an enormous amount of data that requires analysis and interpretation in order to be useful. To facilitate this process, researchers use statistical software to generate various types of summary tables and graphical displays to better understand their data. Histograms, normal Q-Q plots, detrended Q-Q plots, and box plots are common graphs that are used to assess normality of data and identify outliers (i.e., extreme values) in the data. Such graphs are quite useful for identifying data anomalies that might require more in-depth study. Researchers also generate special graphs to ensure that important assumptions are not being violated when performing certain statistical tests (e.g., correlation, *t* test, analysis of variance [ANOVA]). For example, the residuals scatterplot and normal probability plot are useful for checking the assumptions of normality, linearity, and homoscedasticity, as well as identifying outliers.

Narrative and numerical data—no matter how well organized—are of little use if they fail to communicate information. However, many research papers and reports can be intimidating to the average person. Therefore, researchers need to find creative ways to present data so that they are sufficiently appealing to the average reader. Data that are presented in the form of charts and graphs are one way that researchers can make data more appealing; many people find graphs much easier to understand compared to narratives and tables. Effective graphical displays of data can undoubtedly simplify complex data, making it more comprehensible to the average reader. As the old clichè goes—a picture paints a thousand words.

## Common Graphical Displays for Reporting

### *Bar Chart*

Bar charts are one of the most commonly used techniques for presenting data and are considered to be one of the easiest diagrams to read and interpret. They are used to display frequency distributions for categorical variables. In bar chart displays, the value of the observation is proportional to the length of the bar; each category of the variable is represented by a separate bar; and the categories of the variable are generally shown along the horizontal axis, whereas the number of each category is shown on the vertical axis. Bar charts are quite versatile; they can be adapted to incorporate displays of both negative and positive data on the same chart (e.g., profits and losses across years). They are particularly useful for comparing groups and for showing changes over time. Bar charts should generally not contain more than 8–10 categories or they will become cluttered and difficult to read. When more than 10 categories are involved in data analysis, rotated bar charts or line graphs should be considered instead.

### *Pie Chart*

A pie chart is a circle divided into sectors or slices, where the sectors of the pie are proportional to the whole. The entire pie represents 100%. Pie charts are used to display categorical data for a single variable. They are quite popular in journalistic and business reporting. However, these charts can be difficult to interpret unless percentages and/or other numerical information for each slice are shown on the diagram. A good pie chart should have no more than eight sectors or it will become too crowded. One solution is to group several smaller slices into a category called "Other." When color is being used, red and green should not be located on adjacent slices, because some people are color-blind and cannot distinguish red from green. When patterns are used, it is important to ensure that optical illusions are not created on adjacent slices or the data may be misinterpreted.

### *Line Graph*

A line graph shows the relationship between two variables by connecting the data points on a grid, with a line moving from left to right, on the diagram. When several time series lines are being plotted, and color is not being used, pronounced symbols along the lines can help to draw attention to the different variables. For example, a diamond (♦) can be used to represent all the data points for unemployment, a square (■) for job approval, and so on. Another option is to use solid/dotted/dashed lines to distinguish different variables.

## Effective Graphical Displays

The advent of commercial, feature-rich statistical and graphical software such as Excel and IBM® SPSS® (PASW) 18.0 has made the incorporation of professional graphical displays into reports easy and inexpensive. (Note: IBM® SPSS® Statistics was formerly called PASW® Statistics.) Both Excel and SPSS have built-in features that can generate a wide array of graphical displays in mere seconds, using a few point-and-click operations. However, commercial software has also created new problems. For example, some researchers may go overboard and incorporate so many charts and graphs into their writing that the sheer volume of diagrams can make comprehension of the data torturous—rather than enlightening—for the reader. Others may use so many fancy features (e.g., glow, shadows) and design shapes (e.g., cones, doughnuts, radars, cylinders) that diagrams lose their effectiveness in conveying certain information and instead become quite tedious to read. Many readers may become so frustrated that they may never complete reading the document.

An equally problematic issue pertains to distorted and misleading charts and graphs. Some of these distortions may be quite deliberate. For example, sometimes, scales are completely omitted from a graph. In other cases, scales may be started at a number other than zero. Omitting a zero tends to magnify changes. Likewise, "starting time" can also affect the appearance of magnitude. An even worse scenario, however, is when either a "scale" or the "starting time" is adjusted and then combined with a three-dimensional or other fancy graph—this may lead to even greater distortion. Other distortions may simply result from inexperienced persons preparing the diagrams. The resultant effect is that many readers who are not knowledgeable in statistics can be easily misled by such graphs. Thus, when using graphical displays,

meticulous attention should be given to ensuring that the graphs do not emphasize unimportant differences and/or distort or mislead readers.

In order to present effective data, the researcher must be able to identify the salient information from the data. In addition, the researcher must be clear on what needs to be emphasized, as well as the targeted audience for the information. The data must then be presented in a manner that is vivid, clear, and concise. The ultimate goal of effective graphical displays should be to ensure that any data communicated are intelligible and enlightening to the targeted audience. When readers have to spend a great deal of time trying to decipher a diagram, this is a clear indication that the diagram is ineffective.

All graphical displays should include source information. When graphical displays are sourced entirely from other works, written permission is required that will specify exactly how the source should be acknowledged. If graphs are prepared using data that are not considered proprietary, copyright permission need not be sought, but the data source must still be acknowledged. When graphical displays are prepared entirely from the researcher's own data, the source information generally makes reference to the technique/population used to obtain the data (e.g., 2009 Survey of ABC College Students). Source information should be placed at the bottom of the diagram and should be sufficiently detailed to enable the reader to go directly to the source (e.g., Source: General Motors Annual Report, 2009, Page 10, Figure 6—reprinted with permission).

When using graphics, many researchers often concentrate their efforts on ensuring that the salient facts are presented, while downplaying appearance of displays. Others emphasize appearance over content. Both are important. Eye-catching graphs are useless if they contain little or no useful information. On the other hand, a graph that contains really useful content may get limited reading because of its appearance. Therefore, researchers need to package their reports in a manner that would be appealing to a wider mass. Effective data graphics require a combination of good statistical and graphical design skills, which some researchers may not possess. However, numerous guidelines are available on the Internet and in texts that can assist even a novice to create effective graphs. In addition, the following guidelines can assist in creating informative and effective graphical displays that communicate meaningful information with clarity and precision:

1. Focus on substance—emphasize the important.
2. Ensure that data are coherent, clear, and accurate.
3. Use an appropriate scale that will not distort or mislead.
4. Label the *x*-axis and *y*-axis with appropriate labels to aid interpretation [e.g., Temperature (°C); Time (minutes)].
5. Number the graphs/charts, and give them an informative title (e.g., Figure 1: ABC College Course Enrollment 2009).
6. Include the source at the bottom of the diagram.
7. Simplicity is often best. Use three-dimensional and other fancy graphs cautiously—they often distort and/or mislead.
8. Avoid stacked bar charts unless the primary comparison is being made on the data series located on the bottom of the bar.
9. When names are displayed on a label (e.g., countries, universities, etc.), alphabetize data before charting to aid reading.
10. Use statistical and textual descriptions appropriately to aid data interpretation.
11. Use a legend when charts include more than one data series. Locate legend carefully to avoid reducing plot area.
12. Appearance is important. Consider using borders with curved edges and three-dimensional effects to enhance graphical displays. Use colors effectively and consistently. Do not color every graph with a different color. Bear in mind that when colored documents are photocopied in black and white, images will be difficult to interpret unless the original document had sharp color contrast. When original documents are being printed in black and white, it may be best to use shades of black and gray or textual patterns.
13. Avoid chart clutter. This confuses and distracts the reader and can often obscure the distribution's shape. For example, if you are charting data for 20 years, show every other

year on the *x*-axis. Angling labels may create an optical illusion of less clutter.

14. Use readable, clear fonts (e.g., Times New Roman 10 or 12) for labels, titles, scales, symbols, and legends.
15. Ensure that diagrams and legends are not so small that they require a magnifying glass in order to be read.
16. Edit and scale graphs to desired size in the program in which they were created before transferring into the word-processed document to avoid image distortions with resizing.
17. Edit and format graphical displays generated directly from statistical programs before using.
18. Use gridlines cautiously. They may overwhelm and distract if the lines are too thick. However, faded gridlines can be very effective on some types of graphs (e.g., line charts).
19. Use a specific reference format such as APA style to prepare the document to ensure correct placement of graph titles, and so on.
20. Ensure that graphical displays are self-explanatory—readers should be able to understand them with minimal or no reference to the text and tables.

*Nadini Persaud*

***See also*** Bar Chart; Column Graph; Cumulative Frequency Distribution; Histogram; Line Graph; Pie Chart

**Further Readings**

Fink, A. (2003). *How to report on surveys* (2nd ed.). Thousand Oaks, CA: Sage.
Owen, F., & Jones, R. (1994). *Statistics* (4th ed.). London: Pitman.
Pallant, J. (2001). *SPSS survival manual: A step by step guide to using SPSS*. Maidenhead, Berkshire, UK: Open University Press.

# Greenhouse–Geisser Correction

When performing an analysis of variance with a one-factor, repeated-measurement design, the effect of the independent variable is tested by computing an $F$ statistic, which is computed as the ratio of the mean square of effect by the mean square of the interaction between the subject factor and the independent variable. For a design with $S$ subjects and $A$ experimental treatments, when some assumptions are met, the sampling distribution of this $F$ ratio is a Fisher distribution with $\nu_1 = A - 1$ and $\nu_2 = (A - 1)(S - 1)$ degrees of freedom.

In addition to the usual assumptions of normality of the error and homogeneity of variance, the $F$ test for repeated-measurement designs assumes a condition called *sphericity*. Intuitively, this condition indicates that the ranking of the subjects does not change across experimental treatments. This is equivalent to stating that the population correlation (computed from the subjects' scores) between two treatments is the same for all pairs of treatments. This condition implies that there is no interaction between the subject factor and the treatment.

If the sphericity assumption is not valid, then the $F$ test becomes too liberal (i.e., the proportion of rejections of the null hypothesis is larger than the $\alpha$ level when the null hypothesis is true). In order to minimize this problem, Seymour Greenhouse and Samuel Geisser, elaborating on early work by G. E. P. Box, suggested using an index of deviation to sphericity to correct the number of degrees of freedom of the $F$ distribution. This entry first presents this index of nonsphericity (called the Box index, denoted $\varepsilon$), and then it presents its estimation and its application, known as the Greenhouse–Geisser correction. This entry also presents the Huyhn–Feldt correction, which is a more efficient procedure. Finally, this entry explores tests for sphericity.

## Index of Sphericity

Box has suggested a measure for sphericity, denoted $\varepsilon$, which varies between 0 and 1 and reaches the value of 1 when the data are perfectly spherical. The computation of this index is illustrated with the fictitious example given in Table 1 with data collected from $S = 5$ subjects whose responses were measured for $A = 4$ different treatments. The

**Table 1** A Data Set for a Repeated-Measurement Design

| | $a_1$ | $a_2$ | $a_3$ | $a_4$ | | $M_{.s}$ |
|---|---|---|---|---|---|---|
| $S_1$ | 76 | 64 | 34 | 26 | | 50 |
| $S_2$ | 60 | 48 | 46 | 30 | | 46 |
| $S_3$ | 58 | 34 | 32 | 28 | | 38 |
| $S_4$ | 46 | 46 | 32 | 28 | | 38 |
| $S_5$ | 30 | 18 | 36 | 28 | | 28 |
| $M_{a.}$ | 54 | 42 | 36 | 28 | $M_{..} =$ | 40 |

**Table 2** The Covariance Matrix for the Data Set of Table 1

| | $a_1$ | $a_2$ | $a_3$ | $a_4$ | | |
|---|---|---|---|---|---|---|
| $a_1$ | 294 | 258 | 8 | −8 | | |
| $a_2$ | 258 | 94 | 8 | −8 | | |
| $a_3$ | 8 | 8 | 34 | 6 | | |
| $a_4$ | −8 | −8 | 6 | 2 | | |
| $\bar{t}_{a.}$ | 138 | 138 | 14 | −2 | $\bar{t}_{..} =$ | 72 |
| $\bar{t}_{a.} - \bar{t}_{..}$ | 66 | 66 | −8 | −74 | | |

standard analysis of variance of these data gives a value of $F_A = \frac{600}{112} = 5.36$, which, with $\nu_1 = 3$ and $\nu_2 = 12$, has a $p$ value of .014.

In order to evaluate the degree of sphericity, the first step is to create a table called a *covariance matrix*. This matrix is composed of the variances of all treatments and all the covariances between treatments. As an illustration, the covariance matrix for our example is given in Table 2.

Box defined an index of sphericity, denoted $\varepsilon$, which applies to a *population covariance* matrix. If we call $\zeta_{a,a'}$ the entries of this $A \times A$ table, the Box index of nonsphericity is obtained as

$$\varepsilon = \frac{\left(\sum_a \zeta_{a,a}\right)^2}{(A-1)\sum_{a,a'} \zeta^2_{a,a'}} \tag{1}$$

Box also showed that when sphericity fails, the number of degrees of freedom of the $F_A$ ratio depends directly upon the degree of nonsphericity and is equal to $\nu_1 = \varepsilon(A-1)$ and $\nu_2 = \varepsilon(A-1)(S-1)$.

### *Greenhouse–Geisser Correction*

Box's approach works for the *population* covariance matrix, but in general, this matrix is not known. In order to estimate $\varepsilon$, we need to transform the sample covariance matrix into an *estimate* of the population covariance matrix. In order to compute this estimate, we denote by $t_{a,a'}$ the sample estimate of the covariance between groups $a$ and $a'$ (these values are given in Table 2), by $\bar{t}_{a.}$ the mean of the covariances for group $a$, and by $\bar{t}_{..}$ the grand mean of the covariance table. The estimation of the population covariance matrix will have for a general term $s_{a,a'}$, which is computed as

$$\begin{aligned} s_{a,a'} &= (t_{a,a'} - \bar{t}_{..}) - (\bar{t}_{a,.} - \bar{t}_{..}) - (\bar{t}_{a'_{.}} - \bar{t}_{..}) \\ &= t_{a,a'} - \bar{t}_{a,.} - \bar{t}_{a'_{.}} + \bar{t}_{..} \end{aligned} \tag{2}$$

(this procedure is called "double-centering").

Table 3 gives the double-centered covariance matrix. From this matrix, we can compute the estimate of $\varepsilon$, which is denoted $\hat{\varepsilon}$ (compare with Equation 1):

$$\hat{\varepsilon} = \frac{\left(\sum_a s_{a,a}\right)^2}{(A-1)\sum_{a,a'} s^2_{a,a'}}. \tag{3}$$

In our example, this formula gives

$$\begin{aligned} \hat{\varepsilon} &= \frac{(90+90+78+78)^2}{(4-1)(90^2+54^2+\cdots+66^2+78^2)^2} \\ &= \frac{336^2}{3 \times 84,384} = \frac{112,896}{253,152} \end{aligned}$$

We use the value of $\hat{\varepsilon}$ = .4460 to correct the number of degrees of freedom of $F_A$ as $\nu_1 = \hat{\varepsilon}(A-1) = 1.34$

**Table 3** The Double-Centered Covariance Matrix Used to Estimate the Population Covariance Matrix

| | $a_1$ | $a_2$ | $a_3$ | $a_4$ |
|---|---|---|---|---|
| $a_1$ | 90 | 54 | −72 | −72 |
| $a_2$ | 54 | 90 | −72 | −72 |
| $a_3$ | −72 | −72 | 78 | 66 |
| $a_4$ | −72 | −72 | 66 | 78 |

and $\nu_2 = \hat{\varepsilon}(A-1)(S-1) = 5.35$. These corrected values of $\nu_1$ and $\nu_2$ give for $F_A = 5.36$ a probability of $p = .059$. If we want to use the critical value approach, we need to round the values of these corrected degrees of freedom to the nearest integer (which will give here the values of $\nu_1 = 1$ and $\nu_2 = 5$).

### *Eigenvalues*

The Box index of sphericity is best understood in relation to the eigenvalues of a covariance matrix. Covariance matrices belong to the class of positive semidefinite matrices and therefore always have positive or null eigenvalues. Specifically, if we denote by $\mathbf{\Sigma}$ a population covariance, and by $\lambda_\ell$ the $\ell$th eigenvalue of $\mathbf{\Sigma}$, the sphericity condition is equivalent to having all eigenvalues equal to a constant. Formally, the sphericity condition states that

$$\lambda_\ell = \text{constant } \forall \ell. \tag{4}$$

In addition, if we denote by $V$ (also called $\beta$ or $\nu$) the following index,

$$V = \frac{\left(\sum \lambda_\ell\right)^2}{\sum \lambda_\ell^2}, \tag{5}$$

then the Box coefficient can be expressed as

$$\varepsilon = \frac{1}{A-1} V. \tag{6}$$

Under sphericity, all of the eigenvalues are equal, and $V$ is equal to $(A-1)$. The estimate of $\varepsilon$ is obtained by using the eigenvalues of the estimated covariance matrix. For example, the matrix from Table 3 has the following eigenvalues:

$$\lambda_1 = 288, \quad \lambda_2 = 36, \quad \lambda_3 = 12.$$

This gives

$$V = \frac{\left(\sum \lambda_\ell\right)^2}{\sum \lambda_\ell^2} = \frac{(288+36+12)^2}{288^2+36^2+12^2} \approx 1.3379,$$

which, in turn, gives

$$\hat{\varepsilon} = \frac{1}{A-1} V = \frac{1.3379}{3} \approx .4460$$

(this matches the results of Equation 4).

### *Extreme Greenhouse–Geisser Correction*

A *conservative* (i.e., increasing the risk of Type II error: the probability of not rejecting the null hypothesis when it is false) correction for sphericity has been suggested by Greenhouse and Geisser. Their idea is to choose the largest possible value of $\hat{\varepsilon}$, which is equal to $(A-1)$. This leads us to consider that $F_A$ follows a Fisher distribution with $\nu_1 = 1$ and $\nu_2 = S-1$ degrees of freedom. In this case, these corrected values of $\nu_1 = 1$ and $\nu_2 = 4$ give for $F_A = 5.36$ a probability of $p = .081$.

### *Huynh–Feldt Correction*

Huynh Huynh and Leonard S. Feldt suggested a more powerful approximation for $\varepsilon$ denoted $\widetilde{\varepsilon}$ and computed as

$$\widetilde{\varepsilon} = \frac{S(A-1)\hat{\varepsilon} - 2}{(A-1)[S-1-(A-1)\hat{\varepsilon}]}. \tag{7}$$

In our example, this formula gives

$$\widetilde{\varepsilon} = \frac{5(4-1).4460 - 2}{(4-1)[5-1-(4-1).4460]} = .5872.$$

We use the value of $\tilde{\varepsilon} = .5872$ to correct the number of degrees of freedom of $F_A$ as $\nu_1 = \widetilde{\varepsilon}(A-1) = 1.76$ and $\nu_2 = \widetilde{\varepsilon}(A-1)(S-1) = 7.04$. These corrected values give for $F_A = 5.36$ a probability of $p = .041$. If we want to use the critical value approach, we need to round these corrected values for the number of degrees of freedom to the nearest integer (which will give here the values of $\nu_1 = 2$ and $\nu_2 = 7$). In general, the correction of Huynh and Feldt is to be preferred because it is more powerful (and Greenhouse–Geisser is too conservative).

### *Stepwise Strategy for Sphericity*

Greenhouse and Geisser suggest using a stepwise strategy for the implementation of the correction for lack of sphericity. If $F_A$ is not significant with the standard degrees of freedom, there is no need to implement a correction (because it will make it even less significant). If $F_A$ is significant with the extreme correction [i.e., with $\nu_1 = 1$ and $\nu_2 = (S-1)$], then there is no need to correct either (because the correction will make it more significant). If $F_A$

is not significant with the extreme correction but is significant with the standard number of degrees of freedom, then use the $\varepsilon$ correction (they recommend using $\hat{\varepsilon}$, but the subsequent $\tilde{\varepsilon}$ is currently preferred by many statisticians).

## Testing for Sphericity

One incidental question about using a correction for lack of sphericity is to decide when a sample covariance matrix is *not* spherical. Several tests can be used to answer this question. The most well known is Mauchly's test, and the most powerful is the John, Sugiura, and Nagao test.

### *Mauchly's Test*

J. W. Mauchly constructed a test for sphericity based on the following statistic, which uses the eigenvalues of the estimated covariance matrix:

$$W = \frac{\prod \lambda_\ell}{\left[\frac{1}{A-1}\sum \lambda_\ell\right]^{(A-1)}}. \tag{8}$$

This statistic varies between 0 and 1 and reaches 1 when the matrix is spherical. For our example, we find that

$$W = \frac{\prod \lambda_\ell}{\left[\frac{1}{A-1}\sum \lambda_\ell\right]^{(A-1)}} = \frac{228 \times 36 \times 12}{\left[\frac{1}{3}(228+36+12)\right]^3} = \frac{124,416}{1,404,928} \approx .0886$$

Tables for the critical values of $W$ are available in Nagarsenker and Pillai (1973), but a good approximation is obtained by transforming $W$ into

$$X^2_W = -(1-f) \times (S-1) \times \ln\{W\}, \tag{9}$$

where

$$f = \frac{2(A-1)^2 + A + 2}{6(A-1)(S-1)}. \tag{10}$$

Under the null hypothesis of sphericity, $X^2_W$ is approximately distributed as a $\chi^2$ with degrees of freedom equal to

$$\nu = \frac{1}{2}A(A-1). \tag{11}$$

For our example, we find that

$$f = \frac{2(A-1)^2 + A + 2}{6(A-1)(S-1)} = 2 \times 3^2 + 4 + 26 \times 3 \times 4 = \frac{24}{72} = .33$$

and

$$X^2_W = -(1-f) \times (S-1) \times \ln\{W\} = -4(1-.33) \times \ln\{.0886\} \approx 6.46,$$

with $\nu = \frac{1}{2}4 \times 3 = 6$, we find that $p = 38$ and we cannot reject the null hypothesis. Despite its relative popularity, the Mauchly test is not recommended by statisticians because it lacks power. A more powerful alternative is the John, Sugiura, and Nagao test for sphericity described below.

### *John, Sugiura, and Nagao Test*

According to John E. Cornell, Dean M. Young, Samuel L. Seaman, and Roger E. Kirk, the best test for sphericity uses $V$. Tables for the critical values of $W$ are available in A. P. Grieve, but a good approximation is obtained by transforming $V$ into

$$X^2_V = \frac{1}{2}S(A-1)^2\left(V - \frac{1}{A-1}\right). \tag{12}$$

Under the null hypothesis, $X^2_V$ is approximately distributed as a $\chi^2$ distribution with $\nu = \frac{1}{2}A(A-1) - 1$. For our example, we find that

$$X^2_V = \frac{1}{2}S(A-1)^2\left(V - \frac{1}{A-1}\right) = \frac{5 \times 3^2}{2}\left(1.3379 - \frac{1}{3}\right) = 22.60.$$

With $\nu = \frac{1}{2}4 \times 3 - 1 = 5$, we find that $p = .004$ and we can reject the null hypothesis with the usual test. The discrepancy between the conclusions reached from the two tests for sphericity illustrates the lack of power of Mauchly's test.

*Hervé Abdi*

***See also*** Analysis of Covariance (ANCOVA); Analysis of Variance (ANOVA); Pooled Variance; Post Hoc Analysis; Post Hoc Comparisons; Sphericity

**Further Readings**

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, I: Effect of inequality of variance in the one-way classification. *Annals of Mathematical Statistics, 25,* 290–302.

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, II: Effects of inequality of variance and of correlation between errors in the two-way classification. *Annals of Mathematical Statistics, 25,* 484–498.

Cornell, J. E., Young, D. M., Seaman, S. L., & Kirk, R. E. (1992). Power comparisons of eight tests for sphericity in repeated measures designs. *Journal of Educational Statistics, 17,* 233–249.

Geisser, S., & Greenhouse, S. W. (1958). An extension of Box's result on the use of *F* distribution in multivariate analysis. *Annals of Mathematical Statistics*, 29, 885–891.

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika,* 24, 95–112.

Grieve, A. P. (1984). Tests of sphericity of normal distribution and the analysis of repeated measure designs. *Psychometrika,* 49, 257–267.

Huynh, H., & Feldt, L. S. (1970). Conditions under which mean square ratios in repeated measurement designs have exact *F*-distributions. *Journal of the American Statistical Association,* 65, 1582–1589.

Huynh, H., & Feldt, L. S. (1976). Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of Educational Statistics,* 1, 69–82.

John, S. (1972). The distribution of a statistic used for testing sphericity of normal distributions. *Biometrika, 59,* 169–173.

Mauchly, J. W. (1940). Significance test for sphericity of *n*-variate normal population. *Annals of Mathematical Statistics, 11,* 204–209.

Nagao, H. (1973). On some test criteria for covariance matrix. *Annals of Statistics, 1,* 700–709.

Sugiura, N. (1972). Locally best invariant test for sphericity and the limiting distribution. *Annals of Mathematical Statistics, 43,* 1312–1326.

# Grounded Theory

Grounded theory, a qualitative research method, relies on insight generated from the data. Unlike traditional research that begins from a preconceived framework of logically deduced hypotheses, grounded theory begins inductively by gathering data and posing hypotheses during analysis that can be confirmed or disconfirmed during subsequent data collection. Grounded theory is used to generate a theory about a research topic through the systematic and simultaneous collection and analysis of data. Developed in the 1960s by Barney Glaser and Anselm Strauss within the symbolic interactionist tradition of field studies in sociology and drawing also on principles of factor analysis and qualitative mathematics, it is now used widely in the social sciences; business and organizational studies; and, particularly, nursing.

As an exploratory method, grounded theory is particularly well suited for investigating social processes that have attracted little prior research attention, where the previous research is lacking in breadth and/or depth, or where a new point of view on familiar topics appears promising. The purpose is to understand the relationships among concepts that have been derived from qualitative (and, less often, quantitative) data, in order to explore (and explain) the behavior of persons engaged in any specific kind of activity. By using this method, researchers aim to discover the basic issue or problem for people in particular circumstances, and then explain the basic social process (BSP) through which they deal with that issue. The goal is to develop an explanatory theory from the "ground up" (i.e., the theory is derived inductively from the data).

This entry focuses on the grounded theory research process, including data collection, data analysis, and assessments of the results. In addition, modifications to the theory are also discussed.

## Grounded Theory Research Design

One important characteristic that distinguishes grounded theory (and other qualitative research) is the evolutionary character of the research design. Because researchers want to fully understand the meaning and course of action of an experience from the perspective of the participants, variables cannot be identified in advance. Instead, the important concepts emerge during data collection and analysis, and the researcher must remain open-minded to recognize these concepts. Therefore, the research process must be flexible to allow

these new insights to guide further data collection and exploration. At the same time, grounded theory is both a rigorous and systematic approach to empirical research.

### *Writing Memos*

To ensure that a study is both systematic and flexible, the researcher is responsible for keeping detailed notes in the form of memos in which the researcher documents observations in the field, methodological ideas and arrangements, analytical thinking and decisions, and personal reflections. Memo writing begins at the time of conceptualization of the study with the identification of the phenomenon of interest and continues throughout the study. These memos become part of the study data. When a researcher persists in meticulously recording memos, writing the first draft of the study report becomes a simple matter of sorting the memos into a logical sequence.

### *Reviewing the Literature*

Whether to review the literature before data collection may depend on the circumstances of the individual researcher. Methodological purists follow the originators' advice to delay reading related literature to avoid developing preconceived ideas that could be imposed during data analysis, thus ensuring that the conceptualization emerges from the data. Instead, they recommend reading broadly in other disciplines early in the study to develop "sensitizing concepts" that may trigger useful ideas and analogies during the latter stages of theoretical construction and elaboration. For them, the actual literature review is more appropriately begun once the theory has started to take shape, at which time previous writing about those concepts that has already emerged from the data can be helpful for developing theoretical relationships and relating the emerging theory to previous knowledge about the topic. Others, however, recognize pragmatically that for a research proposal to be approved by current funding agencies and thesis committees, knowledge of past research must be demonstrated and then followed up with further reading as the theory develops in order to show where it is congruent (or not) with previous academic work.

### *Data Collection*

Grounded theory is a form of naturalistic inquiry. Because the problems that generate research are located in the natural world, grounded theorists investigate their questions in (and draw their interpretations from) the natural world of their participants. Consequently, once a phenomenon has been identified as the topic to be studied, data collection begins by seeking out the places where the issue occurs and reviewing documents, observing and talking to the people involved, and sometimes reviewing visual media. Consequently, the study begins with purposive sampling. Later, data collection is guided by a particular type of purposive sampling called theoretical sampling.

### *Theoretical Sampling*

After the initial data are analyzed, data collection is directed by the emerging theory. Theoretical sampling is achieved by collecting, coding, and analyzing the data simultaneously, rather than sequentially. Sampling is now guided by deductive reasoning, as the researcher seeks additional data to enlarge upon the insights that have been learned from the participants who have been interviewed thus far and to fill out those portions of the theory that need further development. Because of theoretical sampling, the sample size cannot be determined before the study commences. Only when the researcher is satisfied that no new concepts are emerging and no new information on the important concepts is forthcoming can the decision be made that the point of theoretical saturation has been achieved and data collection ends.

### *Interviewing*

For grounded theory, data consist of any form of information about the research topic that can be gathered, including the researcher's own field notes. For many studies, however, interviews form the majority of the data, but these are usually supplemented with other kinds of information. As a general rule, more than one interview with participants can help to create a broader and more in-depth analysis. Interviewed participants are initially asked broad, open-ended questions to try to elicit their own interpretations and

understandings of what is important in their experience. This relative lack of explicit direction in the questions is a conscious effort not to bias their responses toward what informants might think the researcher wants to hear. As the study evolves, the interview process changes. Later interviews are structured to answer more specific questions aimed at better understanding those concepts that have not yet been fully fleshed out in the data or to seek agreement from participants that the theory accounts for their experience.

Most researchers audiotape the interviews and transcribe them verbatim. This preserves the richness of the data and prepares the data for analysis. Immediately following each interview, written or tape-recorded field notes are made of the interviewer's observations and impressions. Some grounded theorists insist that tape-recording is unnecessary and that a researcher's detailed field notes of conversations are sufficient. They maintain that by relying too heavily on transcriptions, the researcher may be constrained from raising the analysis from a descriptive to a more conceptual and theoretical level of abstraction. After the first or second interview has been completed, the researcher begins constant comparative analysis.

## The Constant Comparative Method

In the analysis of data, grounded theorists employ both inductive and deductive reasoning. Constant comparison is used throughout this simultaneous and iterative collection, analysis, and interpretation of the data. Emerging codes and concepts are continually compared with one another, with new data, with previously analyzed data, and with the researcher's observations and analytical memos.

### *Data Analysis*

Analyzing voluminous amounts of textual data is a matter of data reduction, segmenting the data into sections that can be compared, contrasted, and sorted by categories. Thus, grounded theory analysis consists of coding, categorizing, memo writing, and memo sorting, to arrive at a core variable or BSP that is the central theme of the analysis.

#### Coding and Categorizing Data

Codes and categories are the building blocks of theory. Open coding begins by examining the data minutely, phrase by phrase and line by line, to identify concepts and processes, by asking oneself, "What does this indicate?" and "What is going on here?" These short portions of data are assigned "in vivo" or substantive codes derived as much as possible from the interviewee's own vocabulary. An in-depth interview obviously yields a multitude of codes, although many will be repeated. Codes are repeated when the analyst finds other phrases that indicate the same thing. As the codes are identified, memos are written to define them.

Coding allows the data segments belonging to each code to be sorted together. Comparing these coded segments for similarities and differences allows them to be grouped into categories. As this is done, memos are written for each analytical decision. When each new interview is coded, the substantive codes and the data they contain are compared with other codes and categories, and with the coding in previous interviews. Thus, by comparing incident to incident, data segment with data segment, code with code, and codes with categories and individual cases, connections among the data are identified. Some codes and categories may be combined and the number of conceptual codes reduced. As codes fit together, the relationships among them are recorded in memos.

#### Theoretical Coding

Categories are collapsed into a higher level of theoretical category or construct, as patterns, dimensions, and relationships among them are noted. One way to look for these relationships is to draw diagrams showing connections among the categories and to memos describing those connections. Another is to examine the codes and categories in terms of their causes, contexts, contingencies, consequences, covariances, and conditions (the six Cs). Glaser has described multiple families of theoretical codes, but the six Cs constitute the basic coding family that is commonly employed to tease out the meaning of a code or category. Some other examples are dimensions, degrees, types, and temporal ordering. These coding families are meant to sensitize the analyst to

relationships that may be discovered among codes and categories; they are not intended to serve as a checklist for matching with theoretical constructs.

Theoretical codes or constructs, derived by questioning the data, are used to conceptualize relationships among the codes and categories. Each new level of coding requires the researcher to reexamine the raw data to ensure that they are congruent with the emerging theory. Unanswered questions may identify gaps in the data and are used to guide subsequent interviews until the researcher is no longer able to find new information pertaining to that construct or code. Thus, the code is "saturated," and further data collection omits this category, concentrating on other issues. Throughout, as linkages are discovered and recorded in memos, the analyst posits hypotheses about how the concepts fit together into an integrated theory. Hypotheses are tested against further observations and data collection. The hypotheses are not tested statistically but, instead, through this persistent and methodical process of constant comparison.

#### Hypothesizing a Core Category

Eventually, a core variable that appears to explain the patterns of behavior surrounding the phenomenon of interest becomes evident. This core category links most or all of the other categories and their dimensions and properties together. In most, but not all, grounded theory studies, the core category is a BSP, an "umbrella concept" that appears to explain the essence of the problem for participants and how they attempt to solve it. BSPs may be further subdivided into two types: basic social structural processes and basic social psychological processes.

At this point, however, the core category is only tentative. Further interviews focus on developing and testing this core category by trying to discount it. The researcher presents the theory to new participants and/or previously interviewed participants and elicits their agreement with the theory, further clarification, or refutation. With these new data, the analyst can dispense with open coding and code selectively for the major categories of the BSP. Once satisfied that the theory is saturated and explains the phenomenon, a final literature review is conducted to connect the theory with previous work in the field.

### Substantive and Formal Grounded Theories

Two levels of grounded theory (both of which are considered to be middle-range) can be found in the literature. Most are substantive theories, developed from an empirical study of social interaction in a defined setting (such as health care, education, or an organization) or pertaining to a discrete experience (such as having a particular illness, learning difficult subjects, or supervising co-workers). In contrast, formal theories are more abstract and focused on more conceptual aspects of social interaction, such as stigma, status passage, or negotiation. A common way to build formal theory is by the constant comparative analysis of any group of substantive grounded theories that is focused on a particular social variable but enacted under different circumstances, for different reasons, and in varied settings.

### Using Software for Data Analysis

Using a software program to manage these complex data can expedite the analysis. Qualitative data analysis programs allow the analyst to go beyond the usual coding and categorizing of data that is possible when analyzing the data by hand. Analysts who use manual methods engage in a cumbersome process that may include highlighting data segments with multicolored markers, cutting up transcripts, gluing data segments onto index cards, filing the data segments that pertain to a particular code or category together, and finally sorting and re-sorting these bits of paper by taping them on the walls. Instead, with the aid of computer programs, coding and categorizing the data are accomplished easily, and categorized data segments can be retrieved readily. In addition, any changes to these procedures can be tracked as ideas about the data evolve. With purpose-build programs (such as NVivo), the researcher is also able to build and test theories and construct matrices in order to discover patterns in the data.

Controversy has arisen over the use of computer programs for analyzing qualitative data. Some grounded theorists contend that using a computer program forces the researcher in particular

directions, confining the analysis and stifling creativity. Those who support the use of computers recognize their proficiency for managing large amounts of complex data. Many grounded theorists believe that qualitative software is particularly well-suited to the constant comparative method. Nevertheless, prudent qualitative researchers who use computers as tools to facilitate the examination of their data continually examine their use of technology to enhance, rather than replace, recognized analytical methods.

## Ensuring Rigor

Judging qualitative work by the positivist standards of validity and reliability is inappropriate as these tests are not applicable to the naturalistic paradigm. Instead, a grounded theory is assessed according to four standards, commonly referred to as fit, work, grab, and modifiability.

### *Fit*

To ensure fit, the categories must be generated from the data, rather than the data being forced to comply with preconceived categories. In reality, many of the categories found in the data will be factors that occur commonly in everyday life. However, when such common social variables are found in the data, the researcher must write about these pre-existing categories in a way that reveals their origin in the data. Inserting quotations from the data into the written report is one way of documenting fit.

### *Work*

To work, a theory should explain what happened and variation in how it happened, predict what will happen, and/or interpret what is happening for the people in the setting. Follow-up interviews with selected participants can be used as a check on how well the theory works for them.

### *Grab*

Grab refers to the degree of relevance that the theory and its core concept have to the topic of the study. That is, the theory should be immediately recognizable to participants and others in similar circumstances, as reflective of their own experience.

### *Modifiability*

Finally, modifiability becomes important after the study is completed and when the theory is applied. No grounded theory can be expected to account for changing circumstances. Over time, new variations and conditions that relate to the theory may be discovered, but a good BSP remains applicable because it can be extended and qualified appropriately to accommodate new data and variations.

## Developments in Grounded Theory

In the 50 years that have elapsed since grounded theory was first described in 1967, various grounded theorists have developed modifications to the method. Although Barney Glaser continues to espouse classic grounded theory method, Anselm Strauss and Juliet Corbin introduced the conditional matrix as a tool for helping the analyst to explicate contextual conditions that exert influences upon the action under investigation. Using the conditional matrix model, the analyst is cued to examine the data for the effects of increasingly broad social structures, ranging from groups through organizations, communities, the country, and the international relations within which the action occurs. As new theoretical perspectives came to the fore, grounded theorists adapted the methodology accordingly. For example, Kathy Charmaz contributed a constructivist approach to grounded theory, and Adele Clarke expanded into postmodern thought with situational analysis. Others have used grounded theory within feminist and critical social theory perspectives. Whichever version a researcher chooses for conducting his or her grounded theory study, the basic tenets of grounded theory methodology continue to endure; conceptual theory is generated from the data by way of systematic and simultaneous collection and analysis.

*P. Jane Milliken*

*See also* Inference: Deductive and Inductive; Naturalistic Inquiry; NVivo; Qualitative Research

**Further Readings**

Bryant, A., & Charmaz, K. (2007). *The Sage handbook of grounded theory.* Thousand Oaks, CA: Sage.
Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis.* Thousand Oaks, CA: Sage.
Clarke, A. E. (2005). *Situational analysis: Grounded theory after the postmodern turn.* Thousand Oaks, CA: Sage.
Glaser, B. G. (1978). *Theoretical sensitivity.* Mill Valley, CA: Sociology Press.
Glaser, B. G. (2008). *Doing quantitative grounded theory.* Mill Valley, CA: Sociology Press.
Glaser, B. G., & Strauss, A. (1967). *The discovery of grounded theory.* Chicago: Aldine.
Hutchinson, S. A., & Wilson, H. S. (2001). Grounded theory: The method. In P. L. Munhall (Ed.), *Nursing research: A qualitative perspective* (pp. 209–243). Boston: Jones and Bartlett.
Schreiber, R. S., & Stern, P. N. (2001). *Using grounded theory in nursing.* New York: Springer.
Strauss, A. L. (1987). *Qualitative analysis for social scientists.* Cambridge, UK: Cambridge University Press.
Strauss, A. L., & Corbin, J. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory.* Newbury Park, CA: Sage.

# Group-Sequential Designs in Clinical Trials

If the main interest of a clinical trial is to determine whether a new treatment results in a better outcome than the existing treatment, investigators often would like to obtain the result as soon as possible. One of the reasons for that is that if one treatment is clearly superior to the other, it is unethical to continue the inferior treatment. Standard methods, which fix the length of a study and conduct only one significance test at the end of the study to compare treatments, are inefficient in terms of use of time and cost. Therefore, one question that needs to be answered is whether one can predict with certainty the outcome of the trial before the end of the study based on interim data. Statisticians adapted sequential methods to test a significant difference in treatment groups every time new and follow-up subjects are assessed. Even though this method saves time and requires a smaller number of subjects, it is not adapted to many trials due to its impracticality. As a natural extension of it, the use of group-sequential theory in clinical trials was introduced to accommodate the sequential method's limitations. This entry discusses the group-sequential design methods and describes three procedures for testing significance.

## Methods

Group-sequential methods are clinical trial stopping rules that consist of series of interim analyses conducted at each visit so that any significant difference among treatment groups can be detected before the trial ends. Initially, before the trial begins, the number of visits and the sample size required at each interim visit are determined. Then, at each interim visit, a significance test is conducted. Once there is evidence for significant difference between the treatment groups at any interim visit, an early termination of the clinical trial is possible and there is no need to recruit any more subjects. For the significance testing, there are several available methods, among which Pocock's, O'Brien and Fleming's, and Wang and Tsiatis's tests are widely used in clinical trials.

## Test Procedures

In this section, three tests are described for comparing two treatment groups based on a normal response with known variance. Significance tests for the other types of response variables (e.g., binomial or exponential) are not explained here, but are also available. In addition, if the main interest is to compare more than two treatment groups, it is possible to modify the tests using an $F$ ratio test for one-way analysis of variance. All three tests are similar in the sense that they adjust critical values for multiple comparisons in order to prevent the increasing probability of Type I errors (rejection of a true null hypothesis).

For all tests, let $K$ be the fixed number of visits, which is predetermined before the trial begins. Let $x_{ij}$ be the $j$th subject from the $i$th treatment group, where $i = 1, 2$ and $j = 1, \ldots, n$. Assume that each $x_{ij}$ is independently drawn from a normal distribution with a mean of $\mu_i$ and a variance of $\sigma_i^2$. Finally, $n_k$ is the number of accumulated subjects

at the $k$th visit, and it is assumed that $n_1$ and $n_2$ are even.

### *Pocock's Test*

The method consists of two steps:

1. Calculate $Z_k$ at each visit $k$ where $k = 1, \ldots, K$:

$$Z_k = \frac{1}{\sqrt{n_k(\sigma_1^2 + \sigma_2^2)}} \left( \sum_{j=1}^{n_k} x_{1j} - \sum_{j=1}^{n_k} x_{2j} \right).$$

2. Compare $Z_k$ to a critical value $C_p(K, \alpha)$.

At any interim visit $k$ prior to the final visit $K$, if $|Z_k| > C_p(K, \alpha)$, then stop the trial to conclude that there is evidence that one treatment is superior to the other. Otherwise, continue to collect the assessments. The critical values $C_p(K, \alpha)$ are available in standard textbooks or statistical software packages.

The required sample size per treatment group at each interim visit is calculated as follows:

$$n = R_p(K, \alpha, \beta) \left( \frac{\left(Z_{\alpha/2} + Z_\beta\right)^2 \left(\sigma_1^2 + \sigma_2^2\right)}{\left(\mu_1 - \mu_2\right)^2} \right) / K$$

where $\sigma_1^2$ and $\sigma_2^2$ are the variances of the continuous responses from Treatment Group 1 and 2, respectively. Similarly, $\mu_1$ and $\mu_2$ are the means of the responses from the two treatment groups.

### *O'Brien and Fleming's Test*

This method uses the same $Z_k$ as in Step 1 of Pocock's test. In Step 2, instead of comparing it with $C_p(K, \alpha)$, it is compared with a different critical value $C_B(K, \alpha)\sqrt{K/k}$. Compared to Pocock's test, this test has the advantage of not rejecting the null hypothesis too easily at the beginning of the trial.

The computation of the required sample sizes per treatment group at each interim visit is similar to Pocock's calculation:

$$n = R_B(K, \alpha, \beta) \left( \frac{\left(Z_{\alpha/2} + Z_\beta\right)^2 \left(\sigma_1^2 + \sigma_2^2\right)}{\left(\mu_1 - \mu_2\right)^2} \right) / K$$

### *Wang and Tsiatis's Test*

This method also uses the $Z_k$ from Step 1 in Pocock's test. In Step 2, $Z_k$ is compared with a critical value $C_{WT}(K, \alpha, \Delta)(k/K)^{\Delta - 1/2}$. Refer to *Sample Size Calculations in Clinical Trial Research,* by Shein-Chung Chow, Jun Shao, and Hansheng Wang, for the table of critical values. Pocock's test and O'Brien and Fleming's test are considered to be special cases of Wang and Tsiatis's test.

The calculation of the required sample size per treatment group at each interim visit is formulated as follows:

$$n = R_{WT}(K, \alpha, \beta) \left( \frac{\left(Z_{\alpha/2} + Z_\beta\right)^2 \left(\sigma_1^2 + \sigma_2^2\right)}{\left(\mu_1 - \mu_2\right)^2} \right) / K$$

For calculation of critical values, see the Further Readings section.

*Abdus S. Wahed and Sachiko Miyahara*

***See also*** Cross-Sectional Design; Internal Validity Longitudinal Design; Sequential Design

**Further Readings**

Chow, S., Shao, J., & Wang, H. (2008). *Sample size calculations in clinical trial research*. Boca Raton, FL: Chapman and Hall.

Jennison, C., & Turnbull, B. (2000). *Group sequential methods with applications to clinical trials*. New York: Chapman and Hall.

## Growth Curve

Growth curve analysis refers to the procedures for describing change of an attribute over time and testing related hypotheses. Population growth curve traditionally consists of a graphical display of physical growth (e.g., height and weight) and is typically used by pediatricians to determine whether a specific child seems to be developing as expected. As a research method, the growth curve is particularly useful to analyze and understand longitudinal data. It allows researchers to describe processes that unfold gradually over time for each individual, as well as the differences across individuals, and to systematically relate these differences

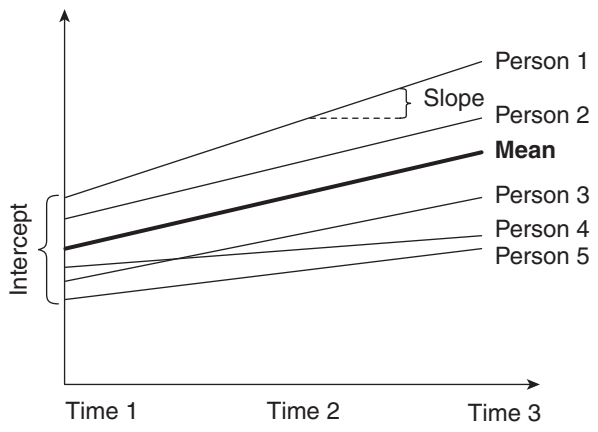



**Figure 1** Individual and Aggregate Growth Curves

against theoretically important time-invariant and time-varying covariates. This entry discusses the use of growth curves in research and two approaches for studying growth curves.

## Growth Curves in Longitudinal Research

One of the primary interests in longitudinal research is to describe patterns of change over time. For example, researchers might be interested in investigating depressive symptoms. Possible questions include the following: Do all people display similar initial levels of depressive symptoms (similar intercepts)? Do some people tend to have a greater increase or decrease in depressive symptoms than others (different slopes)? Separate growth curves can be estimated for each individual, using the following equation:

$$Y_{it} = \beta_{0i} + \beta_{1i}(\text{Time}) + \varepsilon_{it}$$

That is, the outcome variable $Y$ for individual $i$ is predicted by an intercept of $\beta_{0i}$ and a slope of $\beta_{1i}$.The error term at each point in time, $\varepsilon_{it}$, represents the within-subject error. Each individual will have different growth parameters (i.e., different intercept and slope), and these individual growth curves are used to estimate an aggregate mean and variance for the group intercept and the group slope (see Figure 1). The intercept, also called the initial level or constant, represents the value of the outcome variable when the growth curve or change is first measured (when time = 0). The aggregate intercept determines the average outcome variable for all samples, whereas the aggregate slope indicates the average rate of change for the outcome variable for each incremental time point (e.g., year, month, or day).

The growth curve can be positive (an incline) or negative (a decline), linear (representing straight line), or nonlinear. Three or more repeated observations are generally recommended for growth curve analysis. Two waves of data offer very limited information about change and the shape of the growth curves. With three or more waves of data, a linear growth curve can be tested. With four or more waves of data, higher order polynomial alternatives (e.g., quadratic, cubic, logarithmic, or exponential) can be tested. A higher order polynomial growth curve is useful to describe patterns of change that are not the same over time. For example, a rapid increase in weight, height, and muscle mass tends to occur during the first 3 years of childhood; becomes less rapid as children reach their third birthday; and increases rapidly again as they reach puberty. This pattern illustrates the nonlinear trajectories of physical growth that can be captured with additional data points.

In order to measure quantitative changes over time, the study outcome variable must also change continuously and systematically over time. In addition, for each data point, the same instrument must be used to measure the outcome. Consistent measurements help ensure that the changes over time reflect growth and are not due to changes in measurement.

There are two common statistical approaches for studying growth curves. The first approach uses a structural equation modeling framework for estimating growth curves (i.e., latent growth curve analysis). The second approach uses hierarchical linear modeling (i.e., multilevel modeling) framework. These approaches yield equivalent results.

## Growth Curve Within the Structural Equation Modeling Framework

Within the structural equation modeling (SEM) framework, both the initial level and slope are treated as two latent constructs. Repeated

measures at different time points are considered multiple indicators for the latent variable constructs. In a latent growth curve, both the intercept ($\eta_1$) and slope ($\eta_2$) are captured by setting the factor loadings from the latent variable ($\eta$) to the observed variables ($Y$). The intercept is estimated by fixing the factor loadings from the latent variable ($\eta_1$) to the observed variables (repeated measures of $Y_1$ to $Y_3$), each with a value of 1. Latent linear slope loadings are constrained to reflect the appropriate time interval (e.g., 0, 1, and 2 for equally spaced time points). The means for intercept ($\eta_1$) and slope ($\eta_2$) are the estimates of the aggregate intercept and slope for all samples. Individual differences from the aggregate intercept and slope are captured by the variance of the intercept ($\zeta_1$) and slope ($\zeta_2$). The measurement error for each of the three time points is reflected by $\varepsilon_1$, $\varepsilon_2$, and $\varepsilon_3$ (see Figure 2). These measurement errors can be correlated.

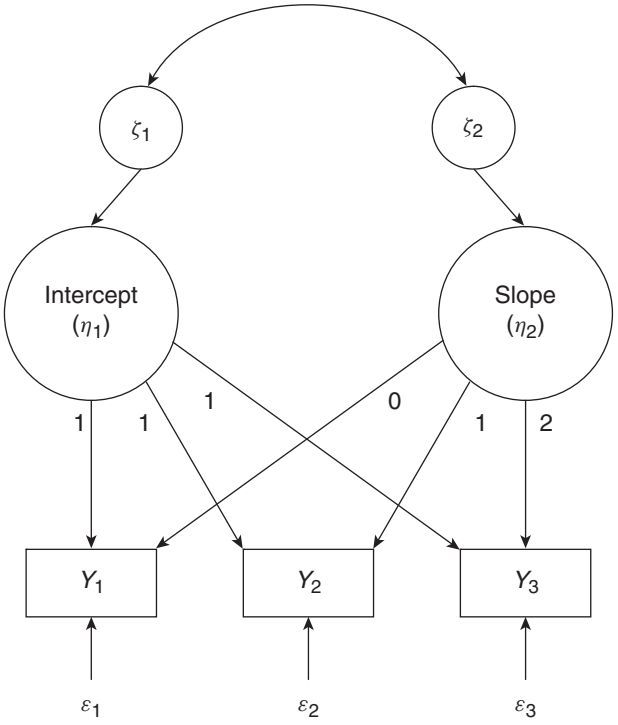



**Figure 2** Univariate Latent Growth Curve in a SEM Framework

## Growth Curve Within the Hierarchical Linear Modeling Framework

In the hierarchical linear modeling (HLM) framework, a basic growth curve model is conceptualized as two levels of analysis. The repeated measures of the outcome of interest are considered "nested" within the individual. Thus, the first level of analysis captures intra-individual changes over time. This is often called the *within-person* level. In the within-person level, just as in the SEM framework, individual growth trajectories are expected to be different from person to person. This approach is more flexible than the traditional ordinary least squares regression technique, which requires the same parameter values for all individuals. The second level of analysis reflects interindividual change and describes between-person variability in the phenomenon of interest.

$$\text{1st level}: Y_{it} = \pi_{0i} + \pi_{1i}(\text{Time})_{it} + e_{it}$$
$$\text{2nd level}: \pi_{0i} = \gamma_{00} + U_{0i}$$
$$\pi_{1i} = \gamma_{10} + U_{1i}$$

$$\text{Combined equations}: Y_{it} = \gamma_{00} + \gamma_{10}(\text{Time})_{it} + U_{0i} + U_{1i} + e_{it}$$

where $\pi_{0i}$ and $\pi_{1i}$ represent the intercept and slope. They are assumed to vary across individuals (as captured by variances $U_{0i}$ and $U_{1i}$). The residuals for each point in time are represented by $e_{it}$ and are assumed to be normally distributed with zero means.

*Florensia F. Surjadi and K. A. S. Wickrama*

***See also*** See also Confirmatory Factor Analysis; Hierarchical Linear Modeling; Hypothesis; Latent Growth Modeling; Latent Variable; Multilevel Modeling; Structural Equation Modeling

### Further Readings

Duncan, T. E., Duncan, S. C., & Strycker, L. A. (2006). *An introduction to latent variable growth curve modeling: Concepts, issues, and applications* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Rogosa, D., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin, 92,* 726–748.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* New York: Oxford University Press.

Wickrama, K. A. S., Beiser, M., & Kaspar, V. (2002). Assessing the longitudinal course of depression and economic integration of South-East Asian refugees: An application of latent growth curve analysis. *International Journal of Methods in Psychiatric Research, 11,* 154–168.

## Guessing Parameter

In item response theory (IRT), the *guessing parameter* is a term informally used for the lower asymptote parameter in a three-parameter-logistic (3PL) model. Among examinees who demonstrate very low levels of the trait or ability measured by the test, the value of the guessing parameter is the expected proportion that will answer the item correctly or endorse the item in the scored direction. This can be understood more easily by examining the 3PL model:

$$P(\theta) = c_i + (1 - c_i)\frac{e^{1.7a_i(\theta - b_i)}}{1 + e^{1.7a_i(\theta - b_i)}}, \qquad (1)$$

where $\theta$ is the value of the trait or ability; P($\theta$) is the probability of correct response or item endorsement, conditional on $\theta$; $a_i$ is the slope or discrimination for item $i$; $b_i$ is the difficulty or threshold for item $i$; and $c_i$ is the lower asymptote or guessing parameter for item $i$. Sometimes, the symbol $g$ is used instead of $c$. In Equation 1, as $\theta$ decreases relative to $b$, the second term approaches zero and thus the probability approaches $c$. If it is reasonable to assume that the proportion of examinees with very low $\theta$ who know the correct answer is virtually zero, it is reasonable to assume that those who respond correctly do so by guessing. Hence, the lower asymptote is often labeled the *guessing parameter.* Figure 1 shows the probabilities from Equation 1 plotted across the range of $\theta$, for $a = 1.5, b = 0.5$, and $c = 0.2$. The range of $\theta$ is infinite, but the range chosen for the plot was $-3$ to $+3$ because most examinees or respondents would fall within this range if the metric were set such that $\theta$ had a mean of 0 and standard deviation of 1 (a common, though arbitrary, way of defining the measurement metric in IRT). This

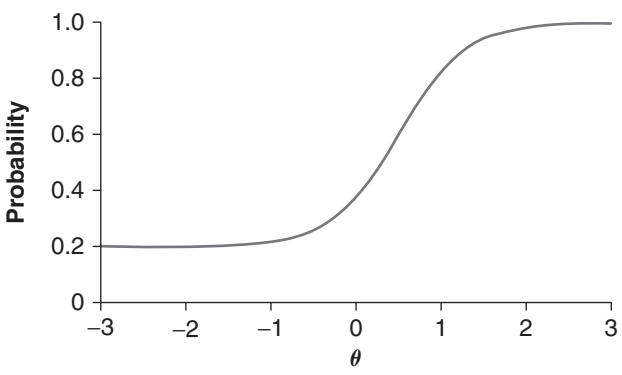



**Figure 1** Item Characteristic Curve

function is called an item characteristic curve (ICC) or item response function (IRF). The value of the lower asymptote or guessing parameter in Figure 1 is 0.2, so as $\theta$ becomes infinitely low, the probability of a correct response approaches 0.2.

Guessing does not necessarily mean random guessing. If the distractors function effectively, the correct answer should be less appealing than the distractors and thus would be selected less than would be expected by random chance. The lower asymptote parameter would then be less than 1/number of options. Frederic Lord suggested that this is often the case empirically for large-scale tests. Such tests tend to be well-developed, with items that perform poorly in pilot testing discarded before the final forms are assembled. In more typical classroom test forms, one or more of the distractors may be implausible or otherwise not function well. Low-ability examinees may guess randomly from a subset of plausible distractors, yielding a lower-asymptote greater than 1/number of options. This could also happen when there is a clue to the right answer, such as the option length. The same effect would occur if examinees can reach the correct answer even by using faulty reasoning or knowledge. Another factor is that examinees who guess tend to choose middle response options; if the correct answer is B or C, the probability of a correct response by guessing would be higher than if the correct answer were A or D.

Because the lower asymptote may be less than or greater than random chance, it may be specified as a parameter to be freely estimated, perhaps with constraints to keep it within

a reasonable range. Estimating the guessing parameter accurately can be difficult. This is particularly true for easy items, because there are few data available at the location where $\theta$ is very low relative to item difficulty. For more difficult items, the value of the guessing parameter makes a bigger difference in the range where the examinee scores are, and an ICC with the wrong guessing parameter does not fit the data well. If the guessing parameter is too low, the best fitting ICC will be too flat and will be too high at one end and too low at the other, especially for more discriminating items. Thus, the guessing parameter can be estimated more accurately for items with high difficulty and discrimination.

Although the term *guessing parameter* is an IRT term, the concept of guessing is also relevant to classical test theory (CTT). For example, when guessing is present in the data, the tetrachoric correlation matrix is often nonpositive definite. Also, CTT scoring procedures can include a guessing penalty to discourage examinees from guessing. For example, one scoring formula is as follows: score $= R - W/(k-1)$, where $R$ is the number of right answers, $W$ is the number of wrong answers, and $k$ is the number of options for each item. If an examinee cannot eliminate any of the options as incorrect and guesses randomly among all options, on average, the examinee would have a probability of $1/k$ of a correct response. For every $k$ items guessed, random guessing would add 1 to $R$ and $k-1$ to $W$. Thus, the average examinee's score would be the same whether the items were left blank or a random guessing strategy was employed. Examinees who could eliminate at least one distractor would obtain a higher score, on average, by guessing among the remaining options than by leaving the item blank. From this, it seems that imposing this penalty for guessing would confound scores with test-taking strategy, an unintended construct. These examples show that guessing is not just a complication introduced by IRT models but a challenge for psychometric modeling in general.

*Christine E. DeMars*

*See also* Classical Test Theory; Item Analysis; Item Response Theory

**Further Readings**

Baker, F. B. (2001). *The basics of item response theory.* College Park, MD: ERIC Clearinghouse on Assessment and Evaluation. Available at http://edres.org/irt

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika, 39,* 247–264.

# Guttman Scaling

Guttman scaling was developed by Louis Guttman and was first used as part of the classic work on the *American Soldier.* Guttman scaling is applied to a set of binary questions answered by a set of subjects. The goal of the analysis is to derive a single dimension that can be used to position both the questions and the subjects. The position of the questions and subjects on the dimension can then be used to give them a numerical value. Guttman scaling is used in social psychology and in education.

## An Example of a Perfect Guttman Scale

Suppose that we test a set of children and that we assess their mastery of the following types of mathematical concepts: (a) counting from 1 to 50, (b) solving addition problems, (c) solving subtraction problems, (d) solving multiplication problems, and (e) solving division problems.

Some children will be unable to master any of these problems, and these children do not provide information about the problems, so we will not consider them. Some children will master counting but nothing more; some will master addition and we expect them to have mastered counting but no other concepts; some children will master subtraction and we expect them to have mastered counting and addition; some children will master multiplication and we expect them to have mastered subtraction, addition, and counting. Finally, some children will master division and we expect them to have mastered counting, addition, subtraction, and multiplication. What we do *not* expect

**Table 1** The Pattern of Responses of a Perfect Guttman Scale

| | *Problems* | | | | |
|---|---|---|---|---|---|
| *Children* | *Counting* | *Addition* | *Subtraction* | *Multiplication* | *Division* |
| $S_1$ | 1 | 0 | 0 | 0 | 0 |
| $S_2$ | 1 | 1 | 0 | 0 | 0 |
| $S_3$ | 1 | 1 | 1 | 0 | 0 |
| $S_4$ | 1 | 1 | 1 | 1 | 0 |
| $S_5$ | 1 | 1 | 1 | 1 | 1 |

*Note:* A value of 1 means that the child (row) has mastered the type of problem (column); a value of 0 means that the child has not mastered the type of problem.

to find, however, are children, for example, who have mastered division but who have not mastered addition or subtraction or multiplication. So, the set of patterns of responses that we expect to find is well structured and is shown in Table 1. The pattern of data displayed in this table is consistent with the existence of a single dimension of mathematical ability. In this framework, a child has reached a certain level of this mathematical ability and can solve all the problems below this level and none of the problems above this level.

When the data follow the pattern illustrated in Table 1, the rows and the columns of the table both can be represented on a single dimension. The operations will be ordered from the easiest to the hardest, and a child will be positioned on the right of the most difficult type of operation solved. So the data from Table 1 can be represented by the following order:

$$\text{Counting}—S_1—\text{Addition}—S_2—\text{Subtraction}—S_3—\text{Multiplication}—S_4—\text{Division}—S_5 \quad (1)$$

This order can be transformed into a set of numerical values by assigning numbers with equal steps between two contiguous points. For example, this set of numbers can represent the numerical values corresponding to Table 1:

| *Counting* | $S_1$ | *Addition* | $S_2$ | *Subtraction* | $S_3$ | *Multiplication* | $S_4$ | *Division* | $S_5$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |

This scoring scheme implies that the score of an observation (i.e., a row in Table 1) is proportional to the number of nonzero variables (i.e., columns in Table 1) for this row.

The previous quantifying scheme assumes that the differences in difficulty are the same between all pairs of contiguous operations. In real applications, it is likely that these differences are not the same. In this case, a way of estimating the size of the difference between two contiguous operations is to consider that this difference is *inversely* proportional to the number of children who solved a given operation (i.e., an easy operation is solved by a large number of children, a hard one is solved by a small number of children).

### *How to Order the Rows of a Matrix to Find the Scale*

When the Guttman model is valid, there are multiple ways of finding the correct order of the rows and the columns that will give the format of the data as presented in Table 1. The simplest approach is to reorder rows and columns according to their marginal sum. Another theoretically interesting procedure is to use correspondence analysis (which is a type of factor analysis tailored for qualitative data) on the data table; then, the coordinates on the first factor of the analysis will provide the correct ordering of the rows and the columns.

## Imperfect Scale

In practice, it is rare to obtain data that fit a Guttman scaling model perfectly. When the data do not conform to the model, one approach is to relax the unidimensionality assumption and assume that the underlying model involves several dimensions. Then, these dimensions can be obtained and analyzed with multidimensional techniques such as correspondence analysis (which can be seen as a multidimensional generalization of Guttman scaling) or multidimensional scaling. Another approach is to consider that the deviations from the ideal scale are random errors. In this case, the problem is to recover the Guttman scale from noisy data. There are several possible ways to fit a Guttman scale to a set of data. The simplest method (called the Goodenough–Edwards method) is to order the rows and the columns according to their marginal sum. An example of

**Table 2** An Imperfect Guttman Scale

| | *Problems* | | | | | |
|---|---|---|---|---|---|---|
| *Children* | *Counting* | *Addition* | *Subtraction* | *Multiplication* | *Division* | Sum |
| $C_1$ | 1 | 0 | 0 | 0 | 0 | 1 |
| $C_2$ | 1 | 0* | 1* | 0 | 0 | 2 |
| $C_3$ | 1 | 1 | 1 | 0 | 0 | 3 |
| $C_4$ | 1 | 1 | 0* | 1 | 0 | 3 |
| $C_5$ | 1 | 1 | 1 | 1 | 1 | 5 |
| Sum | 5 | 3 | 3 | 2 | 1 | – |

*Notes:* Values with an asterisk (*) are considered errors. Compare with Table 1 showing a perfect scale.

a set of data corresponding to such an imperfect scale is given in Table 2. In this table, the "errors" are indicated with an asterisk (*), and there are three of them. This number of errors can be used to compute a coefficient of reproducibility denoted $C_R$ and defined as

$$C_R = 1 - \frac{\text{Number of errors}}{\text{Number of possible errors}}. \quad (2)$$

The number of possible errors is equal to the number of entries in the data table, which is equal to the product of the numbers of rows and columns of this table. For the data in Table 2, there are three errors out of $5 \times 6 = 30$ possible errors; this gives a value of the coefficient of reproducibility equal to

$$C_R = 1 - \frac{3}{30} = .90. \quad (3)$$

According to Guttman, a scale is acceptable if less than 10% of its entries are erroneous, which is equivalent to a scale being acceptable if the value of its $C_R$ is equal to or larger than .90. In practice, it is often possible to improve the $C_R$ of a scale by eliminating rows or columns that contain a large proportion of errors. However, this practice may also lead to capitalizing on random errors and may give an unduly optimistic view of the actual reproducibility of a scale.

*Hervé Abdi*

***See also*** Canonical Correlation Analysis; Categorical Variable; Correspondence Analysis; Likert Scaling; Principal Components Analysis; Thurstone Scaling

**Further Readings**

Dunn-Rankin, P. (1983). *Scaling methods*. Hillsdale, NJ: Lawrence Erlbaum.

Edwards, A. (1957). *Techniques of attitude scale construction*. Englewood Cliffs, NJ: Prentice-Hall.

Greenacre, M. J. (2007). *Correspondence analysis in practice* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Guest, G. (2000). Using Guttman scaling to rank wealth: Integrating quantitative and qualitative data. *Field Methods,* 12, 346–357.

Guttman, L. A. (1944). A basis for scaling qualitative data. *American Sociological Review, 91,* 139–150.

Guttman, L. A. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. A. Guttman, & E. A. Schuman, *Measurement and prediction: Volume 4 of Studies in Social Psychology in World War II*. Princeton, NJ: Princeton University Press.

McIver, J. P., & Carmines, E. G. (1981). *Unidimensional scaling*. Beverly Hills, CA: Sage.

van der Ven, A. H. G. S. (1980). *Introduction to scaling*. Chichester, UK: Wiley.

Weller, S. C., & Romney, A. K. (1990). *Metric scaling: Correspondence analysis*. Thousand Oaks, CA: Sage.

# ENCYCLOPEDIA OF RESEARCH DESIGN

# Editorial Board


## General Editor

Neil J. Salkind
*University of Kansas*

## Associate Editors

Bruce B. Frey
*University of Kansas*

Donald M. Dougherty
*University of Texas Health Science Center at San Antonio*

## Managing Editors

Kristin Rasmussen Teasdale
*University of Kansas*

Nathalie Hill-Kapturczak
*University of Texas Health Science Center at San Antonio*

# ENCYCLOPEDIA OF RESEARCH DESIGN

VOLUME 2

EDITED BY

NEIL J. SALKIND

*University of Kansas*

SAGE | reference

Los Angeles | London | New Delhi
Singapore | Washington DC

Copyright © 2010 by SAGE Publications, Inc.

All rights reserved. No part of this book may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission in writing from the publisher.

*For information:*

SAGE Publications, Inc.
2455 Teller Road
Thousand Oaks, California 91320
E-mail: order@sagepub.com

SAGE Publications Ltd..
1 Oliver's Yard
55 City Road
London EC1Y 1SP
United Kingdom

SAGE Publications India Pvt. Ltd.
B 1/I 1 Mohan Cooperative Industrial Area
Mathura Road, New Delhi 110 044
India

SAGE Publications Asia-Pacific Pte. Ltd.
33 Pekin Street #02-01
Far East Square
Singapore 048763

Printed in the United States of America.

*Library of Congress Cataloging-in-Publication Data*

Encyclopedia of research design/edited by Neil J. Salkind.
v. cm.
Includes bibliographical references and index.
ISBN 978-1-4129-6127-1 (cloth)
1. Social sciences—Statistical methods—Encyclopedias. 2. Social sciences—Research—Methodology—Encyclopedias.
I. Salkind, Neil J.

HA29.E525 2010
001.403—dc22 2010001779

This book is printed on acid-free paper.

10 11 12 13 14 10 9 8 7 6 5 4 3 2 1

| | |
|---|---|
| *Publisher:* | Rolf A. Janke |
| *Acquisitions Editor:* | Jim Brace-Thompson |
| *Editorial Assistant:* | Michele Thompson |
| *Developmental Editor:* | Carole Maurer |
| *Reference Systems Coordinators:* | Leticia M. Gutierrez, Laura Notton |
| *Production Editor:* | Kate Schroeder |
| *Copy Editors:* | Bonnie Freeman, Liann Lech, Sheree Van Vreede |
| *Typesetter:* | C&M Digitals (P) Ltd. |
| *Proofreaders:* | Kristin Bergstad, Kevin Gleason, Sally Jaskold, Sandy Zilka |
| *Indexer:* | Virgil Diodato |
| *Cover Designer:* | Glenn Vogel |
| *Marketing Manager:* | Amberlyn McKay |

# Contents

## Volume 2

List of Entries *vii*

Entries

| | | | |
|---|---|---|---|
| H | *561* | M | *745* |
| I | *589* | N | *869* |
| J | *655* | O | *949* |
| K | *663* | P | *985* |
| L | *681* | | |

# List of Entries


Abstract
Accuracy in Parameter Estimation
Action Research
Adaptive Designs in Clinical Trials
Adjusted *F* Test. *See* Greenhouse–Geisser Correction
Alternative Hypotheses
American Educational Research Association
American Psychological Association Style
American Statistical Association
Analysis of Covariance (ANCOVA)
Analysis of Variance (ANOVA)
Animal Research
Applied Research
A Priori Monte Carlo Simulation
*Aptitudes and Instructional Methods*
Aptitude-Treatment Interaction
Assent
Association, Measures of
Autocorrelation

Bar Chart
Bartlett's Test
Barycentric Discriminant Analysis
Bayes's Theorem
Behavior Analysis Design
Behrens–Fisher *t′* Statistic
Bernoulli Distribution
Beta
Between-Subjects Design. *See* Single-Subject Design; Within-Subjects Design
Bias
Biased Estimator
Bivariate Regression
Block Design
Bonferroni Procedure
Bootstrapping
Box-and-Whisker Plot
b Parameter

Canonical Correlation Analysis
Case-Only Design
Case Study
Categorical Data Analysis
Categorical Variable
Causal-Comparative Design
Cause and Effect
Ceiling Effect
Central Limit Theorem
Central Tendency, Measures of
Change Scores
Chi-Square Test
Classical Test Theory
Clinical Significance
Clinical Trial
Cluster Sampling
Coefficient Alpha
"Coefficient Alpha and the Internal Structure of Tests"
Coefficient of Concordance
Coefficient of Variation
Coefficients of Correlation, Alienation, and Determination
Cohen's *d* Statistic
Cohen's *f* Statistic
Cohen's Kappa
Cohort Design
Collinearity
Column Graph
Completely Randomized Design
Computerized Adaptive Testing
Concomitant Variable
Concurrent Validity
Confidence Intervals
Confirmatory Factor Analysis
Confounding
Congruence
Construct Validity
Content Analysis
Content Validity
Contrast Analysis
Control Group
Control Variables
Convenience Sampling
"Convergent and Discriminant Validation by the Multitrait–Multimethod Matrix"
Copula Functions
Correction for Attenuation
Correlation
Correspondence Analysis
Correspondence Principle
Covariate
C Parameter. *See* Guessing Parameter
Criterion Problem
Criterion Validity
Criterion Variable
Critical Difference
Critical Theory
Critical Thinking
Critical Value
Cronbach's Alpha. *See* Coefficient Alpha

Crossover Design
Cross-Sectional Design
Cross-Validation
Cumulative Frequency Distribution

Databases
Data Cleaning
Data Mining
Data Snooping
Debriefing
Decision Rule
Declaration of Helsinki
Degrees of Freedom
Delphi Technique
Demographics
Dependent Variable
Descriptive Discriminant Analysis
Descriptive Statistics
Dichotomous Variable
Differential Item Functioning
Directional Hypothesis
Discourse Analysis
Discriminant Analysis
Discussion Section
Dissertation
Distribution
Disturbance Terms
*Doctrine of Chances, The*
Double-Blind Procedure
Dummy Coding
Duncan's Multiple Range Test
Dunnett's Test

Ecological Validity
Effect Coding
Effect Size, Measures of
Endogenous Variables
Error
Error Rates
Estimation
Eta-Squared
Ethics in the Research Process
Ethnography
Evidence-Based Decision Making
Exclusion Criteria
Exogenous Variables
Expected Value
Experience Sampling Method
Experimental Design
Experimenter Expectancy Effect
Exploratory Data Analysis
Exploratory Factor Analysis
Ex Post Facto Study
External Validity

Face Validity
Factorial Design
Factor Loadings
False Positive
Falsifiability
Field Study
File Drawer Problem
Fisher's Least Significant Difference Test
Fixed-Effects Models
Focus Group
Follow-Up
Frequency Distribution
Frequency Table
Friedman Test
*F* Test

Gain Scores, Analysis of
Game Theory
Gauss–Markov Theorem
Generalizability Theory
General Linear Model
Graphical Display of Data
Greenhouse–Geisser Correction
Grounded Theory
Group-Sequential Designs in Clinical Trials
Growth Curve
Guessing Parameter
Guttman Scaling

Hawthorne Effect
Heisenberg Effect
Hierarchical Linear Modeling
Histogram
Holm's Sequential Bonferroni Procedure
Homogeneity of Variance
Homoscedasticity
Honestly Significant Difference (HSD) Test
Hypothesis
Inclusion Criteria
Independent Variable
Inference: Deductive and Inductive
Influence Statistics
Influential Data Points
Informed Consent
Instrumentation
Interaction
Internal Consistency Reliability
Internal Validity
Internet-Based Research Method
Interrater Reliability
Interval Scale
Intervention
Interviewing
Intraclass Correlation
Item Analysis
Item Response Theory
Item-Test Correlation

Jackknife
John Henry Effect

Kolmogorov–Smirnov Test
KR-20
Krippendorff's Alpha
Kruskal–Wallis Test
Kurtosis

L'Abbé Plot
Laboratory Experiments
Last Observation Carried Forward
Latent Growth Modeling
Latent Variable
Latin Square Design
Law of Large Numbers
Least Squares, Methods of
Levels of Measurement
Likelihood Ratio Statistic
Likert Scaling
Line Graph
LISREL
Literature Review
*Logic of Scientific Discovery, The*
Logistic Regression
Loglinear Models
Longitudinal Design

Main Effects
Mann–Whitney *U* Test
Margin of Error
Markov Chains
Matching
Matrix Algebra
Mauchly Test
MBESS
McNemar's Test
Mean
Mean Comparisons
Median
Meta-Analysis
"Meta-Analysis of Psychotherapy Outcome Studies"
Methods Section
Method Variance
Missing Data, Imputation of
Mixed- and Random-Effects Models
Mixed Methods Design
Mixed Model Design
Mode
Models
Monte Carlo Simulation
Mortality
Multilevel Modeling
Multiple Comparison Tests
Multiple Regression
Multiple Treatment Interference
Multitrait–Multimethod Matrix
Multivalued Treatment Effects
Multivariate Analysis of Variance (MANOVA)
Multivariate Normal Distribution

Narrative Research
National Council on Measurement in Education
Natural Experiments
Naturalistic Inquiry
Naturalistic Observation
Nested Factor Design
Network Analysis
Newman–Keuls Test and Tukey Test
Nominal Scale
Nomograms
Nonclassical Experimenter Effects
Nondirectional Hypotheses
Nonexperimental Design
Nonparametric Statistics
*Nonparametric Statistics for the Behavioral Sciences*
Nonprobability Sampling
Nonsignificance
Normal Distribution
Normality Assumption
Normalizing Data
Nuisance Variable
Null Hypothesis
Nuremberg Code
NVivo

Observational Research
Observations
Occam's Razor
Odds
Odds Ratio
Ogive
Omega Squared
Omnibus Tests
One-Tailed Test
"On the Theory of Scales of Measurement"
Order Effects
Ordinal Scale
Orthogonal Comparisons
Outlier
Overfitting

Pairwise Comparisons
Panel Design
Paradigm
Parallel Forms Reliability
Parameters
Parametric Statistics
Partial Correlation
Partial Eta-Squared
Partially Randomized Preference Trial Design
Participants
Path Analysis
Pearson Product-Moment Correlation Coefficient
Percentile Rank
Pie Chart
Pilot Study
Placebo
Placebo Effect
Planning Research
Poisson Distribution
Polychoric Correlation Coefficient
Polynomials
Pooled Variance
Population
Positivism
Post Hoc Analysis
Post Hoc Comparisons
Power
Power Analysis
Pragmatic Study
Precision
Predictive Validity
Predictor Variable
Pre-Experimental Design
Pretest–Posttest Design
Pretest Sensitization
Primary Data Source
Principal Components Analysis
*Probabilistic Models for Some Intelligence and Attainment Tests*
Probability, Laws of
Probability Sampling
"Probable Error of a Mean, The"
Propensity Score Analysis
Proportional Sampling
Proposal
Prospective Study
Protocol
"Psychometric Experiments"
Psychometrics
Purpose Statement
*p* Value

Q Methodology
*Q*-Statistic
Qualitative Research
Quality Effects Model
Quantitative Research
Quasi-Experimental Design
Quetelet's Index
Quota Sampling

R
$R^2$
Radial Plot
Random Assignment
Random-Effects Models
Random Error
Randomization Tests
Randomized Block Design
Random Sampling
Random Selection
Random Variable
Range
Rating
Ratio Scale
Raw Scores
Reactive Arrangements
Recruitment
Regression Artifacts
Regression Coefficient
Regression Discontinuity
Regression to the Mean
Reliability
Repeated Measures Design
Replication
Research
Research Design Principles
Research Hypothesis
Research Question
Residual Plot
Residuals
Response Bias
Response Surface Design
Restriction of Range
Results Section
Retrospective Study
Robust
Root Mean Square Error
Rosenthal Effect
Rubrics

Sample
Sample Size
Sample Size Planning
Sampling
Sampling and Retention of Underrepresented Groups
Sampling Distributions
Sampling Error
SAS
Scatterplot
Scheffé Test
Scientific Method
Secondary Data Source
Selection
Semipartial Correlation Coefficient
Sensitivity
Sensitivity Analysis
Sequence Effects
Sequential Analysis
Sequential Design
"Sequential Tests of Statistical Hypotheses"
Serial Correlation
Shrinkage
Significance, Statistical
Significance Level, Concept of
Significance Level, Interpretation and Construction
Sign Test
Simple Main Effects
Simpson's Paradox
Single-Blind Study
Single-Subject Design
Social Desirability
Software, Free
Spearman–Brown Prophecy Formula
Spearman Rank Order Correlation
Specificity
Sphericity
Split-Half Reliability
Split-Plot Factorial Design
SPSS
Standard Deviation
Standard Error of Estimate
Standard Error of Measurement
Standard Error of the Mean
Standardization
Standardized Score
Statistic
Statistica
Statistical Control
*Statistical Power Analysis for the Behavioral Sciences*
Stepwise Regression
Stratified Sampling
Structural Equation Modeling
Student's *t* Test
Sums of Squares
Survey
Survival Analysis
SYSTAT
Systematic Error
Systematic Sampling

"Technique for the Measurement of Attitudes, A"
*Teoria Statistica Delle Classi e Calcolo Delle Probabilità*
Test
Test–Retest Reliability
Theory
Theory of Attitude Measurement
Think-Aloud Methods
Thought Experiments
Threats to Validity
Thurstone Scaling
Time-Lag Study
Time-Series Study
Time Studies
Treatment(s)
Trend Analysis
Triangulation
Trimmed Mean
Triple-Blind Study
True Experimental Design
True Positive
True Score
*t* Test, Independent Samples
*t* Test, One Sample
*t* Test, Paired Samples
Tukey's Honestly Significant Difference (HSD)
Two-Tailed Test
Type I Error
Type II Error
Type III Error

Unbiased Estimator
Unit of Analysis
U-Shaped Curve

"Validity"
Validity of Measurement
Validity of Research Conclusions

Variability, Measure of
Variable
Variance
Volunteer Bias

Wave
Weber–Fechner Law
Weibull Distribution
Weights
Welch's *t* Test
Wennberg Design
White Noise
Wilcoxon Rank Sum Test
WinPepi
Winsorize
Within-Subjects Design
Yates's Correction
Yates's Notation
Yoked Control Procedure

*z* Distribution
Zelen's Randomized Consent Design
*z* Score
*z* Test

# H

## Hawthorne Effect

The term *Hawthorne effect* refers to the tendency for study participants to change their behavior simply as a result of being observed. Consequently, it is also referred to as the *observer effect.* This tendency undermines the integrity of the conclusions researchers draw regarding relationships between variables. Although the original studies from which this term was coined have drawn criticism, the Hawthorne effect remains an important concept that researchers must consider in designing studies and interpreting their results. Furthermore, these studies were influential in the development of a field of psychology known as industrial/organizational psychology.

### History

The term *Hawthorne effect* was coined as a result of events at Hawthorne Works, a manufacturing company outside of Chicago. Throughout the 1920s and early 1930s, officials at the telephone parts manufacturing plant commissioned a Harvard researcher, Elton Mayo, and his colleagues to complete a series of studies on worker productivity, motivation, and satisfaction. Of particular interest to the company was the effect of lighting on productivity; they conducted several experiments to examine that relationship. For example, in one study, the researchers manipulated the level of lighting in order to see whether there were any changes in productivity among a small group of workers at the plant. Consistent with predictions, the researchers found that brighter lighting resulted in increased productivity. Unexpectedly, hourly output also increased when lighting was subsequently dimmed, even below the baseline (or usual) level. In fact, any manipulation or changes to the work environment resulted in increased output for the workers in the study. Decades later, a researcher named Henry Landsberger reevaluated the data and concluded that worker productivity increased simply as a result of the interest being shown in them rather than as a result of changes in lighting or any of the other aspects of the environment the researchers manipulated. Although the term *Hawthorne effect* was derived from this particular series of studies, the term more generally refers to any behavior change that stems from participants' awareness that someone is interested in them.

### A Modern-Day Application

Research on the Hawthorne effect extends well beyond the manufacturing industry. In fact, the effect applies to any type of research. For instance, a recent study examined the Hawthorne effect among patients undergoing arthroscopic knee surgery. In that study, all participants were provided standard preoperative information and informed consent about the operation. However, half of the participants received additional information regarding the

purpose of the study. Specifically, their informed consent form also indicated that they would be taking part in a research study investigating patient acceptability of the side effects of anesthesia. The researchers then examined postoperative changes in psychological well-being and physical complaints (e.g., nausea, vomiting, and pain) in the two groups. Consistent with the Hawthorne effect, participants who received the additional information indicating that they were part of a research study reported significantly better postoperative psychological and physical well-being than participants who were not informed of the study. Similar to the conclusions drawn at the Hawthorne Works manufacturing plant, researchers in the knee surgery study noted that a positive response accompanied simply knowing that one was being observed as part of research participation.

## Threat to Internal Validity

The Hawthorne effect represents one specific type of *reactivity*. Reactivity refers to the influence that an observer has on the behavior under observation and, in addition to the Hawthorne effect, includes *experimenter effects* (the tendency for participants to change their behavior to meet the expectation of researchers), the *Pygmalion effect* (the tendency of students to change their behavior to meet teacher expectations), and the *Rosenthal effect* (the tendency of individuals to internalize the expectations, whether good or bad, of an authority figure). Any type of reactivity poses a threat to interpretation about the relationships under investigation in a research study, otherwise known as *internal validity*. Broadly speaking, the internal validity of a study is the degree to which changes in outcome can be attributed to something the experimenter intended rather than attributed to uncontrolled factors. For example, consider a study in which a researcher is interested in the effect of having a pet on loneliness among the elderly. Specifically, the researcher hypothesizes that elderly individuals who have a pet will report less loneliness than those who do not. To test that relationship, the researcher randomly assigns the elderly participants to one of two groups: one group receives a pet and the other group does not. The researcher schedules monthly follow-up visits with the research participants to interview them about their current level of loneliness. In interpreting the findings, there are several potential reactivity effects to consider in this design. First, failure to find any differences between the two groups could be attributed to the attention that both groups received from the researcher at the monthly visits—that is, reacting to the attention and knowledge that someone is interested in improving their situation (the Hawthorne effect). Conversely, reactivity also could be applied to finding significant differences between the two groups. For example, the participants who received the pet might report less loneliness in an effort to meet the experimenter's expectations (experimenter effects). These are but two of the many possible reactivity effects to be considered in this study.

In sum, many potential factors can impede accurate interpretation of study findings. Reactivity effects represent one important area of consideration when designing a study. It is in the best interest of the researcher to safeguard against reactivity effects to the best of their ability in order to have a greater degree of confidence in the internal validity of their study.

## How to Reduce Threat to Internal Validity

The Hawthorne effect is perhaps the most challenging threat to internal validity for researchers to control. Although *double-blind* studies (i.e., studies in which neither the research participant nor the experimenter are aware to which intervention they are assigned) control for many threats to internal validity, double-blind research designs do not eliminate the Hawthorne effect. Rather, it just makes the effect equal across groups given that everyone knows they are in a research study and that they are being observed. To help mitigate the Hawthorne effect, some have suggested a special design employing what has been referred to as a Hawthorne control. This type of design includes three groups of participants: the control group who receives no treatment, the experimental group who receives the treatment of interest to the researchers, and the Hawthorne control who receives a treatment that is irrelevant to the outcome of interest to the experimenters. For instance, consider the previous example regarding the effect of having a pet on loneliness in the

elderly. In that example, the control group would not receive a pet, the experimental group would receive a pet, and the Hawthorne control group would receive something not expected to impact loneliness such as a book about pets. Thus, if the outcome for the experimental group is significantly different from the outcome of the Hawthorne control group, one can reasonably argue that the specific experimental manipulation, and not simply the knowledge that one is observed, resulted in the group differences.

### Criticisms

In 2009, two economists from the University of Chicago, Steven Levitt and John List, decided to reexamine the data from the Hawthorne Works plant. In doing so, they concluded that peculiarities in the way the studies were conducted resulted in erroneous interpretations that undermined the magnitude of the Hawthorne effect that was previously reported. For instance, in the original experiments, the lighting was changed on Sundays when the plant was closed. It was noted that worker productivity was highest on Monday and remained high during the first part of the workweek but declined as the workweek ended on Saturday. The increase in productivity on Monday was attributed to the change in lighting the previous day. However, further examination of the data indicated that output always was higher on Mondays relative to the end of the previous workweek, even in the absence of experiments. The economists also were able to explain other observations made during the original investigation by typical variance in work performance unrelated to experimental manipulation. Although there has been some controversy regarding the rigor of the original studies at Hawthorne Works and the subsequent inconclusive findings, most people agree that the Hawthorne effect is a powerful but undesirable effect that must be considered in the design of research studies.

### Birth of Industrial Psychology

Regardless of the criticism mounted against the interpretation of the Hawthorne effect, the Hawthorne studies have had lasting impact on the field of psychology. In particular, the Hawthorne studies formed the foundation for the development of a branch of psychology known as industrial/organizational psychology. This particular branch focuses on maximizing the success of organizations and of groups and individuals within organizations. The outcomes of the Hawthorne Works research led to an emphasis on the impact of leadership styles, employee attitudes, and interpersonal relationships on maximizing productivity, an area known as the human relations movement.

*Lisa M. James and Hoa T. Vo*

***See also*** Experimenter Expectancy Effect; Internal Validity; Rosenthal Effect

#### Further Readings

De Amici, D., Klersy, C., Ramajoli, F., Brustia, L., & Politi, P. (2000). Impact of the Hawthorne effect in a longitudinal clinical study: The case of anesthesia. *Controlled Clinical Trials, 21,* 103–114.

Gillespie, R. (1991). *Manufacturing knowledge: A history of the Hawthorne experiments*. Cambridge, UK: Cambridge University Press.

Mayo, E. (1933). *The human problems of an industrial civilization*. New York: MacMillan.

Roethlisberger, F. J., & Dickson, W. J. (1939). *Management and the worker.* Cambridge, MA: Harvard University Press.

Rosenthal, R., & Jacobson, L. (1968, 1992). *Pygmalion in the classroom: Teacher expectation and pupils' intellectual development*. New York: Irvington.

## Heisenberg Effect

Expressed in the most general terms, the Heisenberg effect refers to those research occasions in which the very act of measurement or observation directly alters the phenomenon under investigation. Although most sciences assume that the properties of an entity can be assessed without changing the nature of that entity with respect to those assessed properties, the idea of the Heisenberg effect suggests that this assumption is often violated. In a sense, to measure or observe instantaneously renders the corresponding measurement or observation obsolete. Because reality is not separable from the observer, the process of doing science contaminates reality. Although this term appears frequently in the social

and behavioral sciences, it is actually misleading. For reasons discussed in this entry, some argue it should more properly be called the *observer effect*. In addition, this effect is examined in relation to other concepts and effects.

## Observer Effect

The observer effect can be found in almost any scientific discipline. A commonplace example is taking the temperature of a liquid. This measurement might occur by inserting a mercury-bulb thermometer into the container and then reading the outcome on the instrument. Yet unless the thermometer has exactly the same temperature as the liquid, this act will alter the liquid's post-measurement temperature. If the thermometer's temperature is warmer, then the liquid will be warmed, but if the thermometer's temperature is cooler, then the liquid will be cooled. Of course, the magnitude of the measurement contamination will depend on the temperature discrepancy between the instrument and the liquid. The contamination also depends on the relative amount of material involved (as well as on the specific heat capacities of the substances). The observer effect of measuring the temperature of saline solution in a small vial is far greater than using the same thermometer to assess the temperature of the Pacific Ocean.

As the last example implies, the observer effect can be negligible and, thus, unimportant. In some cases, it can even be said to be nonexistent. If a straight-edge ruler is used to measure the length of an iron bar, under most conditions, it is unlikely that the bar's length will have been changed. Yet even this statement is contingent on the specific conditions of measurement. For instance, suppose that the goal was to measure *in situ* the length of a bar found deep within a subterranean cave. Because that measurement would require the observer to import artificial light and perhaps even inadvertent heat from the observer's body, the bar's dimension could slightly increase. Perhaps the only natural science in which observer effects are completely absent is astronomy. The astronomer can measure the attributes of a remote stellar object, nebula, or galaxy without any fear of changing the phenomenon. Indeed, as in the case of supernovas, the entity under investigation might no longer exist by the time the photons have traveled the immense number of light years to reach the observer's telescopes, photometers, and spectroscopes.

Observer effects permeate many different kinds of research in the behavioral and social sciences. A famous example in industrial psychology is the Hawthorne effect whereby the mere change in environmental conditions can induce a temporary—and often positive—alteration in performance or behavior. A comparable illustration in educational psychology is the Rosenthal or "teacher-expectancy" effect in which student performance is enhanced in response to a teacher's expectation of improved performance. In fact, it is difficult to conceive of a research topic or method that is immune from observer effects. They might intrude on laboratory experiments, field experiments, interviews, and even "naturalistic" observations—the quotes added because the observations cease to be natural to the extent that they are contaminated by observer effects. In "participant observation" studies, the observer most likely alters the observed phenomena to the very degree that he or she actively participates.

Needless to say, observer effects can seriously undermine the validity of the measurement in the behavioral and social sciences. If the phenomenon reacts to assessment, then the resulting score might not closely reflect the true state of the case at time of measurement. Even so, observer effects are not all equivalent in the magnitude of their interference. On the one hand, participants in laboratory experiments might experience evaluation apprehension that interferes with their performance on some task, but this interference might be both small and constant across experimental conditions. The repercussions are thus minimal. On the other hand, participants might respond to certain cues in the laboratory setting—so-called demand characteristics—by deliberately behaving in a manner consistent with their perception of the experimenter's hypothesis. Such artificial (even if accommodating) behavior can render the findings scientifically useless.

Sometimes researchers can implement procedures in the research design that minimize observer effects. A clear-cut instance are the double-blind trials commonly used in biomedical research. Unlike single-blind trials where only the participant is ignorant of the experimental treatment,

double-blind trials ensure that the experimenter is equally unaware. Neither the experimenter nor the participant knows the treatment condition. Such double-blind trials are especially crucial in avoiding the placebo effect, a contaminant that might include an observer effect as one component. If the experimenter is confident that a particular medicine will cure or ameliorate a patient's ailment or symptoms, that expectation alone can improve the clinical outcomes.

Another instance where investigators endeavor to reduce observer effects is the use of deception in laboratory experiments, particularly in fields like social psychology. If research participants know the study's purpose right from the outset, their behavior will probably not be representative of how they would act otherwise. So the participants are kept ignorant, usually by being deliberately misled. The well-known Milgrim experiment offers a case in point. To obtain valid results, the participants had to be told that (a) the investigator was studying the role of punishment on pair-associate learning, (b) the punishment was being administered using a device that delivered real electric shocks, and (c) the learner who was receiving those shocks was experiencing real pain and was suffering from a heart condition. All three of these assertions were false but largely necessary (with the exception of the very last deception).

A final approach to avoiding observer effects is to use some variety of unobtrusive or nonreactive measures. One example is archival data analysis, such as content analysis and historiometry. When the private letters of suicides are content analyzed, the act of measurement cannot alter the phenomenon under investigation. Likewise, when historiometric techniques are applied to biographical information about eminent scientists, that application leaves no imprint on the individuals being studied.

## Quantum Physics

The inspiration for the term *Heisenberg effect* originated in quantum physics. Early in the 20th century, quantum physicists found that the behavior of subatomic particles departed in significant, even peculiar, ways from the "billiard ball" models that prevailed in classical (Newtonian) physics. Instead of a mechanistic determinism in which future events were totally fixed by the prior distributions and properties of matter and energy, reality became much more unpredictable. Two quantum ideas were especially crucial to the concept of the observer effect.

The first is the idea of superimposition. According to quantum theory, it is possible for an entity to exist in all available quantum states simultaneously. Thus, an electron is not in one particular state but in multiple states described by a probability distribution. Yet when the entity is actually observed, it can only be in one specific state. A classic thought experiment illustrating this phenomenon is known as Schrödinger's cat, a creature placed in a box with poison that would be administered contingent on the state of a subatomic particle. Prior to observation, a cat might be either alive or dead, but once it undergoes direct observation, it must occupy just one of these two states. A minority of quantum theorists have argued that it is the observation itself that causes the superimposed states to collapse suddenly into just a single state. Given this interpretation, the result can be considered an observer effect. In a bizarre way, if the cat ends up dead, then the observer killed it by destroying the superimposition! Nonetheless, the majority of theorists do not accept this view. The very nature of observation or measurement in the micro world of quantum physics cannot have the same meaning as in the macro world of everyday Newtonian physics.

The second concept is closest to the source of the term, namely, the 1927 Heisenberg uncertainty principle. Named after the German physicist Werner Heisenberg, this rule asserts that there is a definite limit to how precisely both the momentum and the position of a given subatomic particle, such as an electron, can simultaneously be measured. The more the precision is increased in the measurement of momentum, the less precise will be the concurrent measurement of that particle's position, and conversely. Stated differently, these two particle attributes have linked probability distributions so that if one distribution is narrowed, the other is widened. In early discussions of the uncertainly principle, it was sometimes argued that this trade-off was the upshot of observation. For instance, to determine the location of an electron requires that it be struck with a photon, but that very collision changes the electron's momentum.

Nevertheless, as in the previous case, most quantum theorists perceive the uncertainty as being inherent in the particle and its entanglement with the environment. Position and momentum in a strict sense are concepts in classical physics that again do not mean the same thing in quantum physics. Indeed, it is not even a measurement issue: The uncertainty principle applies independent of the means by which physicists attempt to assess a particle's properties. There is no way to improve measurement so as to lower the degree of uncertainty below a set limit.

In short, the term *Heisenberg effect* has very little, if any, relation with the Heisenberg uncertainty principle—or for that matter any other idea that its originator contributed to quantum physics. Its usage outside of quantum physics is comparable with that of using Einstein's theory of relativity to justify cultural relativism in the behavioral and social sciences. Behavioral and social scientists are merely borrowing the prestige of physics by adopting an eponym, yet in doing so, they end up forfeiting the very conceptual precision that grants physics more status. For this reason, some argue that it would probably be best if the term *Heisenberg effect* was replaced with the term *observer effect*.

## Related Concepts

The observer effect can be confused with other ideas besides the Heisenberg uncertainty principle. Some of these concepts are closely related, and others are not.

An instance of the former is the phenomenon that can be referred to as the *enlightenment effect*. This occurs when the result of scientific research becomes sufficiently well known that the finding renders itself obsolete. In theory the probability of replicating the Milgrim obedience experiment might decline as increasingly more potential research participants become aware of the results of the original study. Although enlightenment effects could be a positive benefit with respect to social problems, they would be a negative cost from a scientific perspective. Science presumes the accumulation of knowledge, and knowledge cannot accumulate if findings cannot be replicated. Still, it must be recognized that the observer and enlightenment effects are distinct. Where in the former the observer directly affects the participants in the original study, in the latter, it is the original study's results that affect the participants in a later partial or complete replication. Furthermore, for good or ill, there is little evidence that enlightenment effects actually occur. Even the widely publicized Milgrim experiment was successfully replicated many decades later.

An example of a divergent concept is also the most superficially similar: observer bias. This occurs when the characteristics of the observer influence how data are recorded or analyzed. Unlike the observer effect, the observer bias occurs in the researcher rather than in the participant. The first prominent example in the history of science appeared in astronomy. Astronomers observing the exact same event—such as the precise time a star crossed a line in a telescope—would often give consistently divergent readings. Each astronomer had a "personal equation" that added or subtracted some fraction of a second to the correct time (defined as the average of all competent observations). Naturally, if observer bias can occur in such a basic measurement, it can certainly infringe on the more complex assessments that appear in the behavioral and social sciences. Hence, in an observational study of aggressive behavior on the playground, two independent researchers might reliably disagree in what mutually observed acts can be counted as instances of aggression. Prior training of the observers might still not completely remove these personal biases. Even so, to the degree that observer bias does not affect the overt behavior of the children being observed, it cannot be labeled as an observer effect.

*Dean Keith Simonton*

***See also*** Experimenter Expectancy Effect; Hawthorne Effect; Interviewing; Laboratory Experiments; Natural Experiments; Naturalistic Observation; Observational Research; Rosenthal Effect; Validity of Measurement

### Further Readings

Chiesa, M., & Hobbs, S. (2008). Making sense of social research: How useful is the Hawthorne effect? *European Journal of Social Psychology, 38,* 67–74.

Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to

demand characteristics and their implications. *American Psychologist, 17,* 776–783.
Rosenthal, R. (2002). *The Pygmalion effect and its mediating mechanisms.* San Diego, CA: Academic Press.
Sechrest, L. (2000). *Unobtrusive measures.* Washington, DC: American Psychological Association.

# Hierarchical Linear Modeling

Hierarchical linear modeling (HLM, also known as multilevel modeling) is a statistical approach for analyzing hierarchically clustered observations. Observations might be clustered within experimental treatment (e.g., patients within group treatment conditions) or natural groups (e.g., students within classrooms) or within individuals (repeated measures). HLM provides proper parameter estimates and standard errors for clustered data. It also capitalizes on the hierarchical structure of the data, permitting researchers to answer new questions involving the effects of predictors at both group (e.g., class size) and individual (e.g., student ability) levels. Although the focus here is on two-level models with continuous outcome variables, HLM can be extended to other forms of data (e.g., binary variables, counts) with more than two levels of clustering (e.g., student, classroom, and school). The key concepts in HLM are illustrated in this entry using a subsample of a publically accessible data set based on the 1982 High School and Beyond (HS&B) Survey. The partial HS&B data set contains a total of 7,185 students nested within 160 high schools, which is included in the free student version of HLM available from Scientific Software International, Inc. Mathematics achievement (MathAch) will be used as the outcome variable in a succession of increasingly complex models. The results of Models A and B discussed here were reported by Stephen Raudenbush and Anthony Bryk in their HLM text.

## Some Important Submodels

### *Model A: Random-Intercepts Model*

The random-intercepts model is the simplest model in which only group membership (here, schools) affects the level of achievement. In HLM, separate sets of regression equations are written at the individual (level 1) and group (level 2) levels of analysis.

**Table 1** Parameter Estimates for Different Models Based on the Full Information Maximum Likelihood (FIML) Estimation

| | *Model A* | *Model B* | *Model C* |
|---|---|---|---|
| *Fixed-effect estimate* | | | |
| Intercept ($\gamma_{00}$) | 12.64 | 12.67 | 13.15 |
| ($SE^{\dagger}$) | (.24) | (.19) | (.21) |
| $SES_{Cij}$ ($\gamma_{10}$) | | 2.40 | 2.55 |
| ($SE^{\dagger}$) | | (.12) | (.14) |
| $Himinty_j$ ($\gamma_{01}$) | | — | −1.86 |
| ($SE^{\dagger}$) | | | (.40) |
| $SES_{Cij}$*$Himinty_j$ ($\gamma_{11}$) | | — | −.57 |
| ($SE^{\dagger}$) | | | (.25) |
| *Variance estimate of the random effect* | | | |
| $\tau_{00}$ | 8.55 | 4.79 | 4.17 |
| $\tau_{11}$ | | .40 | .34 |
| $\tau_{10}$ (or $\tau_{01}$) | | $-.15^{ns}$ | $-.35^{ns}$ |
| $\sigma^2$ | 39.15 | 36.83 | 36.82 |
| Deviance statistic ($D$) | 47,113.97 | 46,634.63 | 46,609.06 |
| Number of parameters | 3 | 6 | 8 |

*Notes:* $^{\dagger}$Standard error of fixed effects in parentheses. All parameter estimates are significant at $p < .05$ unless marked $^{ns}$. Model A contains no predictor. Model B has $SES_{Cij}$ as predictor, and Model C has both $SES_{Cij}$ and $Himinty_j$ as predictors.

**Level 1 (Student-Level) Model**

$$\text{MathAch}_{ij} = \beta_{0j} + e_{ij}, \quad (1)$$

where $i$ represents each student and $j$ represents each school. Note that no predictors are included in Equation 1. $\beta_{0j}$ is the mean MathAch score for school $j$. $e_{ij}$ is the within-school residual that captures the difference between individual MathAch score and the school mean MathAch. $e_{ij}$ is assumed to be normally distributed, and the variance of $e_{ij}$ is assumed to be homogeneous across schools [i.e., $e_{ij} \sim N(0, \sigma^2)$ for all 160 schools]. As presented in Table 1 (Model A), the variance of $e_{ij}$ is equal to $\sigma^2 = 39.15$.

**Level 2 (School-Level) Model**

The level 2 model partitions each school's mean MathAch score into two parts

$$\beta_{0j} = \gamma_{00} + U_{0j}. \quad (2)$$

Here, $\gamma_{00} = 12.64$ is the overall mean MathAch score, averaging over the 160 school means. $U_{0j}$ captures the residual difference between individual school mean MathAch and the overall mean MathAch. $\tau_{00} = \text{Var}(U_{0j}) = 8.55$ is the variance of the residuals at level 2.

The random-intercept model not only provides an important baseline for model comparison, but it also allows the computation of the intraclass correlation (ICC), which is the proportion of the between variance to the sum of the between and within variance.

$$\text{ICC} = \frac{\tau_{00}}{\tau_{00} + \sigma^2}. \quad (3)$$

In this example, the ICC is

$$\text{ICC} = \frac{\tau_{00}}{\tau_{00} + \sigma^2} = \frac{8.55}{8.55 + 39.15} = .18.$$

This ICC generally ranges from 0 to 1 based on Equation 3, with higher values indicating greater clustering. As the product of the ICC and average cluster size increases, the Type I error rate for the study quickly increases to unacceptable levels if ignoring the clustering and treating all observations as independent from each other. For example, with an average of approximately 45 students per school (i.e., 7,185 students/160 schools) and an ICC of approximately .20, a *t* test that compared the mean of two different school types (i.e., high minority enrollment vs. low minority enrollment), but ignored the clustering, would have a Type I error rate of more than .50 compared with the nominal level of $\alpha = .05$. Similarly, the estimated 95% confidence interval is approximately .33 of the width of the correct confidence interval. In contrast, use of HLM procedures can maintain the Type I error rate at the nominal $\alpha = .05$ level.

### *Model B: Random-Coefficients Regression Model With Level 1 Predictor*

The student's socioeconomic status (SES) is added into the model as a level 1 predictor. SES has been centered, $\text{SES}_C = \text{SES} - \text{Mean(SES)}$, so that $\text{SES}_C$ has a mean of 0 in the full sample of 7,185 students.

**Level 1 (Student-Level) Model**

$$\text{MathAch}_{ij} = \beta_{0j} + \beta_{1j}SES_{Cij} + e_{ij}. \quad (4)$$

$\beta_{0j}$ in Equation 4 is the estimated math achievement score in school $j$ for the students who have a mean SES score (i.e., $\text{SES}_C = .00$). $\beta_{1j}$ is the amount of change in MathAch in school $j$ for a 1-unit change in $\text{SES}_C$. $e_{ij}$ is the within-school random error with variance equal to $\sigma^2$ [i.e., $e_{ij} \sim N(0, \sigma^2)$]. Conceptually, this same regression model is applied to each of the 160 schools, yielding 160 different sets of regression coefficients $(\beta_{0j}, \beta_{1j})$.

**Level 2 (School-Level) Model**

$$\beta_{0j} = \gamma_{00} + U_{0j}. \quad (5)$$

$$\beta_{1j} = \gamma_{10} + U_{1j}. \quad (6)$$

As shown in Table 1, Model B, $\gamma_{00} = 12.67$ is the grand intercept, which is the overall average MathAch score for students who have the mean SES score (i.e., $\text{SES}_C = .00$). $\gamma_{10} = 2.40$ is the overall grand slope (or regression coefficient) between SES and MathAch, indicating that MathAch increases 2.40 points for each 1-unit change in SES over the 160 schools, a positive relationship. $U_{0j}$ is the residual from the grand intercept in Equation 5, and $U_{1j}$ is the residual from the grand slope in Equation 6, respectively. The residuals are assumed to have a multivariate normal distribution (i.e., $\begin{bmatrix} U_{0j} \\ U_{1j} \end{bmatrix} \sim N(\underline{0}, T)$, where $T = \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{bmatrix}$), so they can be summarized by three parameters: $\tau_{00} = 4.79$ is the variance of the intercepts, $\tau_{11} = .40$ is the variance of the slopes, and $\tau_{10} = \tau_{01} = -.15$ is the covariance of the slope and intercept across the 160 schools. The negative sign of $\tau_{10}$ $(= \tau_{01})$ indicates that schools with higher intercepts tend to be associated with relatively low slopes (i.e.,

weaker relation) between $SES_C$ and MathAch. Each parameter in Table 1, Model B, can be tested using a Wald $z$ test, $z =$ (estimate)/(standard error). The variances of the intercepts and slopes, $\tau_{00}$ and $\tau_{11}$, are both statistically significant ($p < .05$), whereas the covariance, $\tau_{01}$, between the intercepts and slopes is not ($p > .10$). This result suggests that school-related variables might potentially be able to account for the variation in the regression models across schools.

Raudenbush and Bryk suggest using the baseline random-intercepts model to calculate a measure of the *explained variance* (or pseudo-$R^2$), which is analogous to the $R$-square in the ordinary least-squares (OLS) regression. For the level 1 model, the explained variance is the proportional reduction in the random error variance ($\sigma^2$) of the full model, including all level 1 predictors (here Model B) relative to the random-intercepts model with no predictors (here Model A). Suppose that the estimated random error variance [i.e., $\text{Var}(e_{ij})$] of Model A is $\hat{\sigma}_A$ and that the estimated error variance of the Model B is $\hat{\sigma}_B$. Then, the explained variance [resulting from the additional level 1 predictor(s)] can be calculated by the following equation:

$$\hat{\sigma}^2_{\text{explained}} = \frac{\hat{\sigma}^2_A - \hat{\sigma}^2_B}{\sigma^2_A}. \tag{7}$$

In this example, adding $SES_C$ to the level 1 model can explain

$$\hat{\sigma}^2_{\text{explained}} = \frac{\hat{\sigma}^2_A - \hat{\sigma}^2_B}{\hat{\sigma}^2_A} = \frac{39.15 - 36.83}{39.15} = .06,$$

.06 or 6% of the variance in MathAch.

### *Model C: Random-Coefficients Regression Model With Level 1 and Level 2 Predictors*

A second predictor, high minority percentage (i.e., Himinty), is added to the level 2 model in addition to $SES_C$ at level 1. Himinty is a dummy variable in which 44 schools with greater than 40% minority enrollment are coded as 1 and 116 schools with less than 40% minority enrollment are coded as 0. The level 1 equation (Equation 4) does not change, and the corresponding level 2 models are presented as follows.

**Level 2 (School-Level) Model**

$$\beta_{0j} = \gamma_{00} + \gamma_{01}\text{Himinty}_j + U_{0j}, \tag{8}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}\text{Himinty}_j + U_{1j}. \tag{9}$$

To interpret the meaning of the fixed-effect coefficients in Equations 8 and 9, it is useful to substitute in the corresponding values of the dummy codes for different type of school—0 or 1.

$$\textit{Low minority school} : \beta_{0j} = \gamma_{00} + \gamma_{01}(0) + U_{0j}\beta_{0j} = \gamma_{00} + \gamma_{01}(1) + U_{0j}$$

$$\textit{High minority school} : \beta_{1j} = \gamma_{10} + \gamma_{11}(0) + U_{1j}\beta_{1j} = \gamma_{10} + \gamma_{11}(1) + U_{1j}$$

As shown in Table 1, Model C, $\gamma_{00} = 13.15$ is the mean intercept, which is the mean MathAch score for students with mean SES (i.e., $SES_C = .00$) in schools with relatively low minority enrollment. $\gamma_{10} = 2.55$ is the mean slope of the relation between MathAch and SES for the schools with the low percentage of minority students. $\gamma_{01} = -1.86$ is the difference in the mean intercepts between the two types of schools, which is the difference in mean MathAch scores for students with mean SES ($SES_C = 0$). $\gamma_{11} = -.57$ is the difference in mean slopes between the two types of schools. All fixed-effect estimates/regression coefficients are significant, and the corresponding models for the two school types are presented in Figure 1. On average, high minority-school students with mean SES score ($SES_C = 0$) are lower on MathAch (by $\gamma_{01} = -1.86$ points) than their low minority-school counterparts. The average slope between MathAch and SES is weaker for the high minority-school students ($\gamma_{10} + \gamma_{11} = 2.55 - .57 = 1.98$ MathAch points per 1-unit increase in $SES_C$) than the low minority-school students ($\gamma_{10} = 2.55$ MathAch points per 1-unit increase in $SES_C$).

The explained level 2 variance for each of the two random effects can also be calculated using the explained variance as described previously. Here, Model B is the unconditional model (without any level 2 predictors) and Model C is the

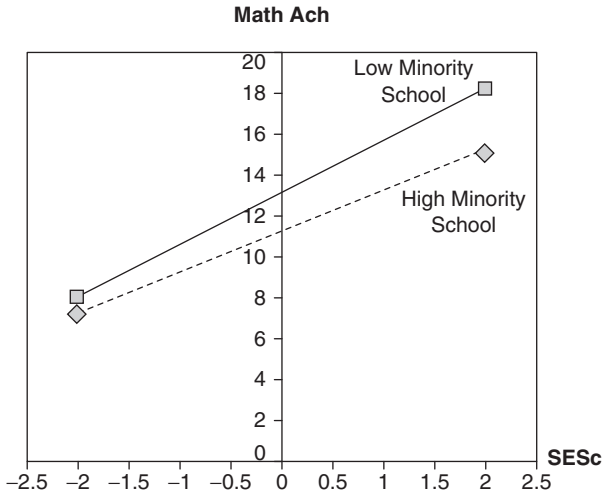



**Figure 1** The Estimated Overall Average Models for the Two Different Types of Schools

conditional model (with Himinty in the model), and the corresponding explained variances for the two random effect variances are

$$\tau_{00_\text{explained}} = \frac{\tau_{00_B} - \tau_{00_C}}{\tau_{00_B}} = \frac{4.79 - 4.17}{4.79} = .13$$

and

$$\tau_{11_\text{explained}} = \frac{\tau_{11_B} - \tau_{11_C}}{\tau_{11_B}} = \frac{.40 - .34}{.40} = .15.$$

That is, .13 (or 13%) of the intercept variance ($\tau_{00}$) and .15 (or 15%) of the slope variance ($\tau_{11}$) can be explained by adding the school type variable (i.e., Himinty) in the level 2 model. Note that although this way of calculating explained variance is straightforward, it is *not* fully analogous to $R^2$ in multiple regression—it can sometimes result in a negative explained variance.

## Other Important Features

### *Estimation Methods and Model Comparison*

Restricted maximum likelihood (REML), full information maximum likelihood (FIML), and empirical Bayesian (EB) estimation are commonly used in HLM to provide accurate parameter estimation. Among these three common estimation methods, REML is preferable for estimating the fixed-effect parameters in HLM, especially when the number of level 2 units/clusters is small.

Deviance statistics can be used to compare nested models. If Model 1 is a special case of Model 2 (i.e., Model 2 can be changed into Model 1 by imposing constraints), then Model 1 is nested within Model 2. In this example, Model B is nested within Model C because Model B is a special case of Model C in which both $\gamma_{01}$ and $\gamma_{11}$ are constrained to zero. A higher deviance indicates poorer fit. The difference in the deviance statistics between nested models follows a central chi-square distribution with degrees of freedom equal to the difference in the number of parameters between the models. The REML deviance statistic can be used only for models with nested random effects. In contrast, models with both nested random and fixed effects can be compared using the FIML deviance statistics. For example, one can examine whether the addition of Himinty to both the intercept and slope equations contributes significantly to the overall model by comparing the deviance statistics between Models B and C, which are estimated with FIML. The significant difference in the deviance statistics (i.e., $\chi^2(2) = D_B - D_C = 46{,}634.63 - 46{,}609.06 = 25.57$, $p < .001$) indicates that the more complex model (i.e., Model C with Himinty) fits the data better than the parsimonious model (i.e., Model B with fewer parameters). However, a nonsignificant difference in the deviance statistics indicates that the fit of the parsimonious model to the data does not differ from that of the complex model. Additionally, criteria such as Akaike information criterion (AIC) and Bayes information criterion (BIC) combine information from the deviance statistic, model complexity, and sample size to help select the model with the optimal combination of fit and parsimony.

### *Examining Assumptions*

Violations of the assumptions including normality and homoscedasticity of the level 1 and level 2 residuals in HLM can be treated as a signal of misspecification in the hypothesized model. The homogeneity assumption can be examined through the probability plot of the standardized residual dispersions, and the normality assumption can be examined through a q−q plot in which the ordered expected Mahalanobis distances are

plotted against the observed Mahalanobis distances across all clusters.

### *Diagnostics*

Toby Lewis and colleague proposed a top-down procedure (i.e., from highest level to lowest level) using diagnostic measures such as leverage, internally and externally Studentized residuals, and DFFITS for different level observations, which are analogous to diagnostic measures commonly used in OLS regression.

*Oi-Man Kwok, Stephen G. West, and Ehri Ryu*

***See also*** Fixed-Effects Models; Intraclass Correlation; Multilevel Modeling; Multiple Regression; Random-Effects Models

### Further Readings

de Leeuw, J., & Meijer, E. (2008). *Handbook of multilevel analysis*. New York: Springer.
Hox, J. J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Lawrence Erlbaum.
Kreft, I. G. G., & de Leeuw, J. (1998). *Introducing multilevel modeling*. Thousand Oaks, CA: Sage.
Lewis, T., & Langford, I. H. (2001). Outliers, robustness and the detection of discrepant data. In A. H. Leyland & H. Goldstein (Eds.), *Multilevel modelling of health statistics* (pp. 75–91). New York: Wiley.
O'Connell, A. A., & McCoach, D. B. (2008). *Multilevel modeling of educational data*. Charlotte, NC: Information Age Publishing.
Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.
Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage.

### Websites

Scientific Software International, Inc. HLM 6, Student Version: http://www.ssicentral.com/hlm/student.html

# HISTOGRAM

A histogram is a method that uses bars to display count or frequency data. The independent variable consists of interval- or ratio-level data and is usually displayed on the abscissa ($x$-axis), and the frequency data on the ordinate ($y$-axis), with the height of the bar proportional to the count. If the data for the independent variable are put into "bins" (e.g., ages 0–4, 5–9, 10–14, etc.), then the width of the bar is proportional to the width of the bin. Most often, the bins are of equal size, but this is not a requirement. A histogram differs from a bar chart in two ways. First, the independent variable in a bar chart consists of either nominal (i.e., named, unordered categories, such as religious affiliation) or ordinal (ranks or ordered categories, such as stage of cancer) data. Second, to emphasize the fact that the independent variable is not continuous, the bars in a bar chart are separated from one another, whereas they abut each other in a histogram. After a bit of history, this entry describes how to create a histogram and then discusses alternatives to histograms.

## A Bit of History

The term *histogram* was first used by Karl Pearson in 1895, but even then, he referred to it as a "common form of graphical representation," implying that the technique itself was considerably older. Bar charts (along with pie charts and line graphs) were introduced over a century earlier by William Playfair, but he did not seem to have used histograms in his books.

## Creating a Histogram

Consider the hypothetical data in Table 1, which tabulates the number of hours of television watched each week by 100 respondents. What is immediately obvious is that it is impossible to comprehend what is going on. The first step in trying to make sense of these data is to put them in rank order, from lowest to highest. This says that the lowest value is 0 and the highest is 64, but it does not yield much more in terms of understanding. Plotting the raw data would result in several problems. First, many of the bars will have heights of zero (e.g., nobody reported watching for one, two, or three hours a week), and most of the other bars will be only one or two units high (i.e., the number of people reporting that specific value). This leads to the second problem, in that it makes

**Table 1** Fictitious Data on How Many Hours of Television Are Watched Each Week by 100 People

| *Subjects* | *Data* | | | | |
|---|---|---|---|---|---|
| 1–5 | 41 | 43 | 14 | 35 | 31 |
| 6–10 | 39 | 22 | 9 | 32 | 49 |
| 11–15 | 12 | 27 | 53 | 7 | 23 |
| 16–20 | 29 | 22 | 22 | 26 | 14 |
| 21–25 | 33 | 34 | 12 | 13 | 16 |
| 26–30 | 34 | 25 | 40 | 5 | 41 |
| 31–35 | 43 | 30 | 40 | 44 | 12 |
| 36–40 | 55 | 14 | 25 | 32 | 10 |
| 41–45 | 30 | 28 | 25 | 23 | 0 |
| 46–50 | 56 | 24 | 17 | 15 | 33 |
| 51–55 | 30 | 15 | 29 | 20 | 14 |
| 56–60 | 40 | 26 | 24 | 34 | 49 |
| 61–65 | 50 | 26 | 13 | 36 | 47 |
| 66–70 | 19 | 9 | 64 | 35 | 33 |
| 71–75 | 35 | 39 | 9 | 25 | 41 |
| 76–80 | 5 | 18 | 54 | 11 | 59 |
| 81–85 | 36 | 36 | 37 | 52 | 29 |
| 86–90 | 24 | 22 | 41 | 36 | 31 |
| 91–95 | 32 | 10 | 50 | 45 | 23 |
| 96–100 | 24 | 15 | 5 | 20 | 52 |

**Table 2** The Data in Table 1 Grouped into Bins

| *Interval* | *Midpoint* | *Count* | *Cumulative total* |
|---|---|---|---|
| 0–4 | 2 | 1 | 1 |
| 5–9 | 7 | 7 | 8 |
| 10–14 | 12 | 12 | 20 |
| 15–19 | 17 | 7 | 27 |
| 20–24 | 22 | 13 | 40 |
| 25–29 | 27 | 12 | 52 |
| 30–34 | 32 | 14 | 66 |
| 35–39 | 37 | 10 | 76 |
| 40–44 | 42 | 10 | 86 |
| 45–49 | 47 | 4 | 90 |
| 50–54 | 52 | 6 | 96 |
| 55–59 | 57 | 3 | 99 |
| 60–64 | 62 | 1 | 100 |

it difficult to discern any pattern. Finally, the $x$-axis will have many values, again interfering with comprehension.

The solution is to group the data into mutually exclusive and collectively exhaustive classes, or bins. The issue is how many bins to use. Most often, the answer is somewhere between 6 and 15, with the actual number depending on two considerations. The first is that the bin size should be an easily comprehended size. Thus, bin sizes of 2, 5, 10, or 20 units are recommended, whereas those of 3, 7, or 9 are not. The second consideration is esthetics; the graph should get the point across and not look too cluttered.

Although several formulas have been proposed to determine the width of the bins, the simplest is arguably the most useful. It is the range of the values (largest minus smallest) divided by the desired number of bins. For these data, the range is 64, and if 10 bins are desired, it would lead to a bin width of 6 or 7. Because these are not widths that are easy to comprehend, the closest compromise would be 5. Table 2 shows the results of putting the data in bins of five units each. The first column lists the values included in each bin; the second column provides the midpoint of the bin; and the third column summarizes the number of people in each bin. The last column, which gives the cumulative total for each interval, is a useful check that the counting was accurate.

However, there is a price to pay for putting the data into bins, and it is that some information is lost. For example, Table 2 shows that seven people watched between 15 and 19 hours of television per week, but the exact amounts are now no longer known. In theory, all seven watched 17 hours each. In reality, only one person watched 17 hours, although the mean of 17 for these people is relatively accurate. The larger the bin width, the more information that is lost.

The scale on the $y$-axis should allow the largest number in any of the bins to be shown, but again it should result in divisions that are easy to grasp. For example, the highest value is 14, but if this were chosen as the top, then the major tick marks would be at 0, 7, and 14, which is problematic for the viewer. It would be better to extend the $y$-axis to 15, which will result in tick marks every five units, which is ideal. Because the data being plotted are counts or frequencies, the $y$-axis most often starts at zero. Putting all this together results in the histogram in Figure 1.

The exception to the rule of the $y$-axis starting at zero is when all of the bars are near the top of the graph. In this situation, small but important differences might be hidden. When this occurs, the

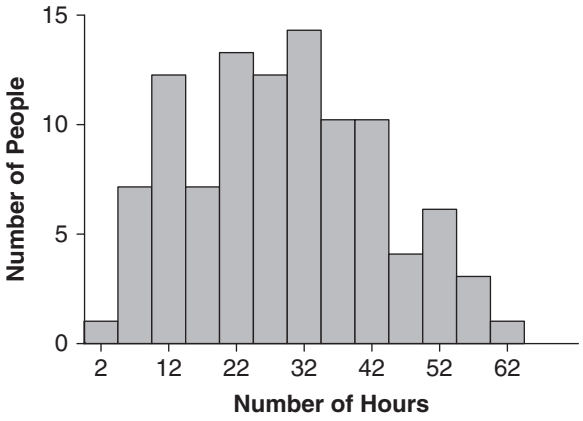



**Figure 1** Histogram Based on Table 2

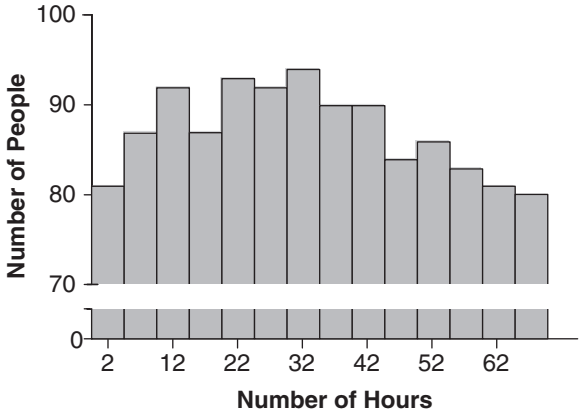



**Figure 2** Histogram Showing a Discontinuity in the *y*-Axis

bottom value should still be zero, and there would be a discontinuity before the next value. It is important, though, to flag this for the viewer, by having a break within the graph itself, as in Figure 2.

The histogram is an excellent way of displaying several attributes about a distribution. The first is its shape—is it more or less normal, or rectangular, or does it seem to follow a power function, with counts changing markedly at the extremes? The second attribute is its symmetry—is the distribution symmetrical, or are there very long or heavy tails at one end? This is often seen if there is a natural barrier at one end, beyond which the data cannot go, but no barrier at the other end. For example, in plotting length of hospitalization or time to react to some stimulus, the barrier at the left is zero for days in hospital (a logical limit) and approximately 50 milliseconds for reaction time (a physiological limit), with no upper limit. Such an asymmetry could be a warning not to use certain statistical tests with these data, if the tests assume that the data are normally distributed. Finally, the graph can easily show whether the data are unimodal, bimodal, or have more than two peaks that stand out against all of the other data.

### Alternatives to Histograms

One alternative to a histogram is a frequency polygon. Instead of drawing bars, a single point is placed at the top of the bin, corresponding to its midpoint, and the points are connected with lines. The only difference between a histogram and a frequency polygon is that, by convention, an extra bin is placed at the upper and lower ends with a frequency of zero, tying the line to the *x*-axis. Needless to say, this is omitted if that category is nonsensical (e.g., an age less than 0).

Another variation is the stem-and-leaf display. This is a histogram where the "bars" consist of the actual values, combining a graph with a data table. Its advantage is that no information is lost because of binning.

*David L. Streiner*

*See also* Graphical Display of Data; Line Graph; Pie Chart

#### Further Readings

Cleveland, W. S. (1985). *The elements of graphing data.* Pacific Grove, CA: Wadsworth.

Robbins, N. B. (2005). *Creating more effective graphs.* Hoboken, NJ: Wiley.

Streiner, D. L., & Norman, G. R. (2007). *Biostatistics: The bare essentials* (3rd ed.). Shelton, CT: People's Medical Publishing House.

## Holm's Sequential Bonferroni Procedure

The more statistical tests one performs the more likely one is to reject the null hypothesis when it is

true (i.e., a false alarm, also called a Type 1 error). This is a consequence of the logic of hypothesis testing: The null hypothesis for *rare* events is rejected in this entry, and the larger the number of tests, the easier it is to find rare events that are false alarms. This problem is called the *inflation* of the alpha level. To be protected from it, one strategy is to correct the alpha level when performing multiple tests. Making the alpha level more stringent (i.e., smaller) will create less errors, but it might also make it harder to detect real effects. The most well-known correction is called the Bonferroni correction; it consists in multiplying each probability by the total number of tests performed. A more powerful (i.e., more likely to detect an effect exists) sequential version was proposed by Sture Holm in 1979. In Holm's sequential version, the tests need first to be performed in order to obtain their $p$ values. The tests are then ordered from the one with the smallest $p$ value to the one with the largest $p$ value. The test with the lowest probability is tested first with a Bonferroni correction involving all tests. The second test is tested with a Bonferroni correction involving one less test and so on for the remaining tests. Holm's approach is more powerful than the Bonferroni approach, but it still keeps under control the inflation of the Type 1 error.

## The Different Meanings of Alpha

When a researcher performs more than one statistical test, he or she needs to distinguish between two interpretations of the $\alpha$ level, which represents the probability of a Type 1 error. The first interpretation evaluates the probability of a Type 1 error for the whole set of tests, whereas the second evaluates the probability for only one test at a time.

### *Probability in the Family*

A *family of tests* is the technical term for a series of tests performed on a set of data. This section shows how to compute the probability of rejecting the null hypothesis at least once in a family of tests when the null hypothesis is true.

For convenience, suppose that the significance level is set at $\alpha = .05$. For each test the probability of making a *Type I error* is equal to $\alpha = .05$. The events "making a Type I error" and "not making a Type I error" are *complementary events* (they cannot occur simultaneously). Therefore, the probability of *not making a Type I error* on one trial is equal to

$$1 - \alpha = 1 - .05 = .95.$$

Recall that when two events are *independent*, the probability of observing these two events together is the *product* of their probabilities. Thus, if the tests are independent, the probability of not making a Type I error on the first *and* the second tests is

$$.95 \times .95 = (1 - .05)^2 = (1 - \alpha)^2.$$

With three tests, the probability of not making a Type I error on all tests is

$$.95 \times .95 \times .95 = (1 - .05)^3 = (1 - \alpha)^3.$$

For a family of $C$ tests, the probability of *not* making a Type I error for the *whole family* is

$$(1 - \alpha)^C.$$

For this example, the probability of not making a Type I error on the family is

$$(1 - \alpha)^C = (1 - .05)^{10} = .599.$$

Now, the probability of making one or more Type I errors on the family of tests can be determined. This event is the complement of the event *not making a Type I error on the family,* and therefore, it is equal to

$$1 - (1 - \alpha)^C.$$

For this example,

$$1 - (1 - .05)^{10} = .401.$$

So, with an $\alpha$ level of .05 for *each* of the 10 tests, the probability of incorrectly rejecting the null hypothesis is .401.

This example makes clear the need to distinguish between two meanings of $\alpha$ when performing multiple tests:

1. The probability of making a Type I error when dealing only with a specific test. This probability is denoted $\alpha[PT]$ (pronounced "alpha *per* test"). It is also called the *testwise* alpha.

2. The probability of making at least one Type I error for the whole family of tests. This probability is denoted $\alpha[PT]$ (pronounced "alpha *per* family of tests"). It is also called the *familywise* or the *experimentwise* alpha.

### *How to Correct for Multiple Tests*

Recall that the probability of making *at least one* Type I error for a family of C tests is

$$\alpha[PF] = 1 - (1 - \alpha[PT])^{C}. \quad (1)$$

This equation can be rewritten as

$$\alpha[PT] = 1 - (1 - \alpha[PF])^{1/C}. \quad (2)$$

This formula—derived assuming *independence* of the tests—is sometimes called the Šidák equation. It shows that in order to maintain a given $\alpha[PF]$ level, the $\alpha[PT]$ values used for each test need to be adapted.

Because the Šidák equation involves a fractional power, it is difficult to compute by hand, and therefore, several authors derived a simpler approximation, which is known as the *Bonferroni* (the most popular name), *Boole,* or even *Dunn* approximation. Technically, it is the first (linear) term of a Taylor expansion of the Šidák equation. This approximation gives

$$\alpha[PF] \approx C \times \alpha[PT] \quad (3)$$

and

$$\alpha[PT] \approx \frac{\alpha[PF]}{C}. \quad (4)$$

Šidák and Bonferroni are linked to each other by the inequality

$$\alpha[PT] = 1 - (1 - \alpha[PF])^{1/C} \geq \frac{\alpha[PF]}{C}. \quad (5)$$

They are, in general, very close to each other, but the Bonferroni approximation is pessimistic (it always does worse than the Šidák equation). Probably because it is easier to compute, the Bonferroni approximation is more well known (and cited more often) than the exact Šidák equation.

The Šidák–Bonferroni equations can be used to find the value of $\alpha[PT]$ when $\alpha[PF]$ is fixed. For example, suppose that you want to perform four *independent* tests, and you want to limit the risk of making at least one Type I error to an overall value of $\alpha[PF] = .05$, you will consider a test significant if its associated probability is smaller than

$$\alpha[PT] = 1 - (1 - \alpha[PF])^{1/C} = 1 - (1 - .05)^{1/4} = .0127.$$

With the Bonferroni approximation, a test reaches significance if its associated probability is smaller than

$$\alpha[PT] = \frac{\alpha[PF]}{C} = \frac{.05}{4} = .0125,$$

which is very close to the exact value of .0127.

### *Bonferroni and Šidák Correction for a* p *Value*

When a test has been performed as part of a family comprising C tests, the $p$ value of this test can be corrected with the Šidák or Bonferroni approaches by replacing $\alpha[PF]$ by $p$ in Equations 1 or 3. Specifically, the Šidák corrected $p$ value for C comparisons, denoted $p_{\text{Šidák,C}}$ becomes

$$p_{\text{Šidák, C}} = 1 - (1 - p)^{C}, \quad (6)$$

and the Bonferroni corrected $p$ value for C comparisons, denoted $p_{\text{Bonferroni, C}}$ becomes

$$p_{\text{Bonferroni, C}} = C \times p\ . \quad (7)$$

Note that the Bonferroni correction can give a value of $p_{\text{Šidák, C}}$ larger than 1. In such cases, $p_{\text{Bonferroni, C}}$ is set to 1.

## Sequential Holm–Šidák and Holm–Bonferroni

Holm's procedure is a sequential approach whose goal is to increase the power of the statistical tests while keeping under control the familywise Type I error. As stated, suppose that a family comprising C tests is evaluated. The first step in Holm's procedure is to perform the tests to obtain their $p$ values, Then the tests are ordered from the one with the smallest $p$ value to the one with the largest $p$ value. The test with the smallest probability will be tested with a Bonferroni or a Šidák correction for a family of C tests (Holm used a Bonferroni correction, but Šidák gives an accurate value

and should be preferred to Bonferroni, which is an approximation). If the test is not significant, then the procedure stops. If the first test is significant, the test with the second smallest $p$ value is then corrected with a Bonferroni or a Šidák approach for a family of $(C - 1)$ tests. The procedure stops when the first nonsignificant test is obtained or when all the tests have been performed. Formally, assume that the tests are ordered (according to their $p$ values) from 1 to $C$, and that the procedure stops at the first nonsignificant test. When using the Šidák correction with Holm's approach, the corrected $p$ value for the $i$th test, denoted $p_{\text{Šidák, } i/C}$, is computed as

$$p_{\text{Šidák, } i|C} = 1 - (1 - p)^{C-i+1}. \tag{8}$$

When using the Bonferroni correction with Holm's approach, the corrected $p$ value for the $i$th test, denoted $p_{\text{Bonferroni, } i/C}$, is computed as

$$p_{\text{Bonferroni, } i|C} = (C - i + 1) \times p. \tag{9}$$

Just like the standard Bonferroni procedure, corrected $p$ values larger than 1 are set equal to 1.

### *Example*

Suppose that a study involving analysis of variance has been designed and that there are three tests to be performed. The $p$ values for these three tests are equal to 0.000040, 0.016100, and 0.612300 (they have been ordered from the smallest to the largest). Thus, $C = 3$. The first test has an original $p$ value of $p = 0.000040$. Because it is the first of the series, $i = 1$, and its corrected $p$ value using the Holm–Šidák approach (cf. Equation 8) is equal to

$$\begin{aligned} p_{\text{Šidák, } i|C} &= 1 - (1 - p)^{C-i+1} = p_{\text{Šidák, } 3|3} \\ &= 1 - (1 - 0.000040)^{3-1+1} \\ &= p_{\text{Šidák, } 1|3} = 1 - (1 - 0.000040)^3 \\ &= 1 - 0.99960^3 \\ &= 0.000119\ . \end{aligned} \tag{10}$$

Using the Bonferroni approximation (cf. Equation 9) will give a corrected $p$ value of $p_{\text{Bonferroni, } 1/3} = .000120$. Because the corrected $p$ value for the first test is significant, the second test can then be performed for which $i = 2$ and $p = .016100$. Using Equations 8 and 9, the corrected $p$ values of $p_{\text{Šidák, } 2/3} = .031941$ and $p_{\text{Bonferroni, } 2/3} = .032200$ are found. The corrected $p$ values are significant, and, so, the last test can be performed for which $i = 3$. Because this is the last of the series, the corrected $p$ values are now equal to the uncorrected $p$ value of $p = p_{\text{Šidák, } 3/3} = p_{\text{Bonferroni, } 3/3} = .612300$, which is clearly not significant. Table 1 gives the results of the Holm's sequential procedure along with the values of the standard Šidák and Bonferroni corrections.

## Correction for Nonindependent Tests

The Šidák equation is derived assuming *independence* of the tests. When they are not independent, it gives a conservative estimate. The Bonferonni being a conservative estimation of Šidák will also give a conservative estimate. Similarly, the sequential Holm's approach is conservative when the tests are not independent. Holm's approach is obviously more powerful than Šidák's (because the $p_{\text{Šidák, } i|C}$ values are always smaller than or equal to the $p_{\text{Šidák, } i|C}$ values), but it still controls the overall familywise error rate. The larger the number of tests, the larger the increase in power with Holm's

**Table 1** Šidák, Bonferroni, Holm–Šidák, and Holm–Bonferroni Corrections for Multiple Comparisons for a Set of $C = 3$ Tests with $p$ Values of 0.000040, 0.016100, and 0.612300, Respectively

| | | *Šidák* | *Bonferroni* | *Holm–Šidák* | *Holm–Bonferroni* |
|---|---|---|---|---|---|
| $i$ | $\alpha[PT]$<br>$p$ | $p_{\text{Šidák, } C}$<br>$1 - (1 - p)^{C-i+1}$ | $p_{\text{Bonferroni, } C}$<br>$C \times p$ | $p_{\text{Šidák, } i\|C}$<br>$1 - (1 - p)^{C-i+1}$ | $p_{\text{Bonferroni, } C}$<br>$(C - i + 1) \times p$ |
| 1 | 0.000040 | 0.000119 | 0.000120 | 0.000119 | 0.000120 |
| 2 | 0.016100 | 0.047526 | 0.048300 | 0.031941 | 0.032200 |
| 3 | 0.612300 | 0.941724 | 1.000000 | 0.612300 | 0.612300 |

procedure compared with the standard Šidák (or Bonferroni) correction.

### Alternatives

The Šidák–Bonferroni as well as Holm's approaches become very conservative when the number of comparisons becomes large and when the tests are not independent (e.g., as in brain imaging). Recently, some alternative approaches have been proposed to make the correction less stringent. A more recent approach redefines the problem by replacing the notion of $\alpha[PF]$ by the false discovery rate (FDR), which is defined as the ratio of the number of Type I errors by the number of significant tests.

*Hervé Abdi*

***See also*** Bonferroni Procedure; Post Hoc Analysis; Post Hoc Comparisons; *Teoria Statistica Delle Classi e Calcolo Delle Probabilità*

#### Further Readings


Abdi, H., Edelman, B., Valentin, D., & Dowling, W. J. (2009). *Experimental design and analysis for psychology.* Oxford, UK: Oxford University Press.

Aickin, M., & Gensler, H. (1996). Adjusting for multiple testing when reporting research results: The Bonferroni *vs.* Holm methods. *American Journal of Public Health,* 86, 726–728.

Benjamini, Y., & Hochberg, T. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B,* 57, 289–300.

Games, P. A. (1977). An improved $t$ table for simultaneous control on $g$ contrasts. *Journal of the American Statistical Association,* 72, 531–534.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika,* 75, 800–803.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics,* 6, 65–70.

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology,* 46, 561–584.

Šidák, Z. (1967). Rectangular confidence region for the means of multivariate normal distributions. *Journal of the American Statistical Association,* 62, 626–633.


## Homogeneity of Variance

Homogeneity of variance is an assumption underlying both $t$ tests and $F$ tests (analyses of variance, ANOVAs) in which the population variances (i.e., the distribution, or "spread," of scores around the mean) of two or more samples are considered equal. In correlations and regressions, the term "homogeneity of variance in arrays," also called "homoskedasticity," refers to the assumption that, within the population, the variance of $Y$ for each value of $X$ is constant. This entry focuses on homogeneity of variance as it relates to $t$ tests and ANOVAs.

### Homogeneity Within Populations

Within research, it is assumed that populations under observation (e.g., the population of female college students, the population of stay-at-home fathers, or the population of older adults living with type 2 diabetes) will be relatively similar and, therefore, will provide relatively similar responses or exhibit relatively similar behaviors. If two identifiable samples (or subpopulations) are each extracted from a larger population, the assumption is that the responses, measurable behaviors, and so on, of participants within both groups will be similar and that the distribution of responses measured within each of the groups (i.e., variance) will also be similar. It is important to note, however, that it would be unreasonable to expect that the variances be exactly equal, given fluctuations based on random sampling. When testing for homogeneity of variance, the goal is to determine whether the variances of these groups are relatively similar or different. For example, is the variation in responses of female college students who attend large public universities different from the variation in responses of female college students who attend small private universities? Is the variation in observable behaviors of older adults with type 2 diabetes who exercise different from that of older adults with type 2 diabetes who do not exercise? Is the variation in responses of 40-year-old stay-at-home fathers different from 25-year-old stay-at-home fathers?

## Assumptions of *t* tests and ANOVAs

When the null hypothesis is $H_0 : \mu_1 = \mu_2$, the assumption of homogeneity of variance must first be considered. Note that testing for homogeneity of variance is different from hypothesis testing. In the case of *t* tests and ANOVAs, the existence of statistically significant differences between the means of two or more groups is tested. In tests of homogeneity of variance, differences in the variation of the distributions among subgroups are examined.

The assumption of homogeneity of variance is one of three underlying assumptions of *t* tests and ANOVAs. The first two assumptions concern independence of observations, that scores within a given sample are completely independent of each other (e.g., that an individual participant does not provide more than one score or that participants providing scores are not related in some way), and normality (i.e., that the scores of the population from which a sample is drawn are normally distributed). As stated, homogeneity of variance, the third assumption, is that the population variances ($\sigma^2$) of two or more samples are equal ($\sigma_1^2 = \sigma_2^2$). It is important to remember that the underlying assumptions of *t* tests and ANOVAs concern populations not samples. In running *t* tests and ANOVAs, the variances of each of the groups that have been sampled are used in order to test this assumption.

This assumption, that the variances are equal (or similar), is quite tenable. In experimental research methods, studies often begin with a "treatment" group and a "control" group that are assumed to be equal at the experiment's onset. The treatment group is often manipulated in a way that the researcher hopes will change the measurable behaviors of its members. It is hoped that the participants' scores will be raised or lowered by an amount that is equivalent to the "effect" of the experimental treatment. If it is the treatment effect alone that raises or lowers the scores of the treatment group, then the variability of the scores should remain unchanged (note: the values will change, but the spread or distribution of the scores should not). For example, a researcher might want to conduct a study on the blood sugar levels of older adults with type 2 diabetes and obtain two samples of older adults with type 2 diabetes. This researcher believes that 30 minutes of exercise, 4 times a week, will lower the blood sugar levels of older adults by 5 points within the first two months. The population variance of the two samples would be assumed to be similar at the onset of the study. It is also the case that introducing a constant (in this case, the treatment effect, or exercise) either by addition or by subtraction has no effect on the variance of a sample. If all values of each of the participants in the treatment sample decrease by five points, while the values of each of the participants in the control sample remain the same, the similarities of the variance will not change simply because the values of the treatment participants all decreased by the same value (the treatment effect).

The variances of groups are "heterogeneous" if the homogeneity of variance assumption has been violated. ANOVAs are robust (not overly influenced by small violations of assumptions) even when the homogeneity of variance assumption is violated, if there are relatively equal numbers of subjects within each of the individual groups.

## Pooled Variance

In conducting *t* tests and ANOVAs, the population variance ($\sigma^2$) is estimated using sample data from both (all) groups. The homogeneity of variance assumption is capitalized on in *t* tests and ANOVAs when the estimates of each of the samples are averaged. Based on the multiple groups, a pooled variance estimate of the population is obtained. The homogeneity of variance assumption ($\sigma_1^2 = \sigma_2^2$) is important so that the pooled estimate can be used. The pooling of variances is done because the variances are assumed to be equal and estimating the same quantity (the population variance) in the first place. If sample sizes are equal, the pooling of variances will yield the same result. However, when sample sizes are unequal, the pooling of variances can cause quite different results.

## Testing for Homogeneity of Variance

When testing for homogeneity of variance, the null hypothesis is $H_0 : \sigma_1^2 = \sigma_2^2$. The ratio of the two variances might also be considered. If the two variances are equal, then the ratio of the variances equals 1.00. Therefore, the null hypothesis is

$H_0 : \sigma_1^2/\sigma_2^2 = 1$. When this null hypothesis is not rejected, then homogeneity of variance is confirmed, and the assumption is not violated.

The standard test for determining homogeneity of variance is the Levene's test and is most frequently used in newer versions of statistical software. Alternative approaches to Levene's test have been proposed by O'Brien and by Brown and Forsythe. For a more detailed presentation on calculating the Levene's test by hand, refer to Howell, 2007. Generally, tests of homogeneity of variance are tests on the deviations (squared or absolute) of scores from the sample mean or median. If, for example, Group A's deviations from the mean or median are larger than Group B's deviations, then it can be said that Group A's variance is larger than Group B's. These deviations will be larger (or smaller) if the variance of one of the groups is larger (or smaller). Based on the Levene's test, it can be determined whether a statistically significant difference exists between the variances of two (or more) groups.

If the result of a Levene's test is not statistically significant, then there are no statistical differences between the variances of the groups in question and the homogeneity of variance assumption is met. In this case, one fails to reject the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ that the variances of the populations from which the samples were drawn are the same. That is, the variances of the groups are not statistically different from one another, and $t$ tests and ANOVAs can be performed and interpreted as normal. If the result of a Levene's test is statistically significant, then the null hypothesis, that the groups have equal variances, is rejected. It is concluded that there are statistically significant differences between the variances of the groups and the homogeneity of variance assumption has been violated. Note: The significance level will be determined by the researcher (i.e., whether the significance value exceeds .05, .01, etc.).

When the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ is rejected, and the homogeneity of variance assumption is violated, it is necessary to adjust the statistical procedure used and employ more conservative methods for testing the null hypothesis $H_0 : \mu_1 = \mu_2$. In these more conservative procedures, the standard error of difference is estimated differently and the degrees of freedom that are used to test the null hypothesis are adjusted. However, these more conservative methods, which alleviate the problem heterogeneity of variance, leave the researcher with less statistical power for hypothesis testing to determine whether differences between group means exist. That is, the researcher is less likely to obtain a statistically significant result using the more conservative method.

## Robustness of *t* tests and ANOVAs

Problems develop when the variances of the groups are extremely different from one another (if the value of the largest variance estimate is more than four or five times that of the smallest variance estimate), or when there are large numbers of groups being compared in an ANOVA. Serious violations can lead to inaccurate $p$ values and estimates of effect size. However, $t$ tests and ANOVAs are generally considered robust when it comes to moderate departures from the underlying homogeneity of variance assumption. Particularly when group sizes are equal ($n_1 = n_2$) and large. If the group with the larger sample also has the larger variance estimate, then the results of the hypothesis tests will be too conservative. If the larger group has the smaller variance, then the results of the hypothesis test will be too liberal. Methodologically speaking, if a researcher has violated the homogeneity of variance assumption, he or she might consider equating the sample sizes.

## Follow-Up Tests

If the results of an ANOVA are statistically significant, then post hoc analyses are run to determine where specific group differences lie, and the results of the Levene's test will determine which post hoc tests are run and should be examined. Newer versions of statistical software provide the option of running post hoc analyses that take into consideration whether the homogeneity of variance assumption has been violated.

Although this entry has focused on homogeneity of variance testing for $t$ tests and ANOVAs, tests of homogeneity of variance for more complex statistical models are the subject of current research.

*Cynthia R. Davis*

*See also* Analysis of Variance (ANOVA); Student's *t* Test; *t* Test, Independent Samples; *t* Test, Paired Samples; Variance

### Further Readings

Boneau, C. A. (1960). The effects of violations of assumptions underlying the *t* test. *Psychological Bulletin, 57,* 49–64.

Games, P. A., & Howell, J. F. (1976). Pairwise multiple comparison procedures with unequal *n*'s and/or variances: A Monte Carlo study. *Journal of Educational Statistics, 1,* 113–125.

Howell, D. C. (2007). *Fundamental statistics for the behavioral sciences* (6th ed.). Belmont, CA: Duxbury.

# HOMOSCEDASTICITY

Homoscedasticity suggests equal levels of variability between quantitative dependent variables across a range of independent variables that are either continuous or categorical. This entry focuses on defining and evaluating homoscedasticity in both univariate and multivariate analyses. The entry concludes with a discussion of approaches used to remediate violations of homoscedasticity.

## Homoscedasticity as a Statistical Assumption

Homoscedasticity is one of three major assumptions underlying parametric statistical analyses. In univariate analyses, such as the analysis of variance (ANOVA), with one quantitative dependent variable ($Y$) and one or more categorical independent variables ($X$), the homoscedasticity assumption is known as *homogeneity of variance*. In this context, it is assumed that equal variances of the dependent variable exist across levels of the independent variables.

In multivariate analyses, homoscedasticity means all pairwise combinations of variables ($X$ and $Y$) are normally distributed. In regression contexts, homoscedasticity refers to constant variance of the residuals (i.e., the difference between the actual and the predicted value of a data point), or *conditional variance,* regardless of changes in $X$.

### *Heteroscedasticity*

Violation of the homoscedasticity assumption results in *heteroscedasticity* when values of the dependent variable seem to increase or decrease as a function of the independent variables. Typically, homoscedasticity violations occur when one or more of the variables under investigation are not normally distributed. Sometimes heteroscedasticity might occur from a few discrepant values (*atypical data points*) that might reflect actual extreme observations or recording or measurement error.

Scholars and statisticians have different views on the implications of heteroscedasticity in parametric analyses. Some have argued that heteroscedasticity in ANOVA might not be problematic if there are equal numbers of observations across all cells. More recent research contradicts this view and argues that, even in designs with relatively equal cell sizes, heteroscedasticity increases the Type I error rate (i.e., error of rejecting a correct null hypothesis). Still others have persuasively argued that heteroscedasticity might be substantively interesting to some researchers.

Regardless, homoscedasticity violations result in biased statistical results and inaccurate inferences about the population. Therefore, before conducting parametric analyses, it is critical to evaluate and address normality violations and examine data for outlying observations. Detection of homoscedasticity violations in multivariate analyses is often made post hoc, that is, by examining the variation of residuals values.

## Exploratory Data Analysis

### *Evaluating Normality*

Generally, normality violations for one or more of the variables under consideration can be evaluated and addressed in the early stages of analysis. Researchers suggest examining a few characteristics of single-variable distributions to assess normality. For example, the location (i.e., anchoring point of a distribution that is ordered from the lowest to highest values, often measured by mean, median, or mode) and spread of data (i.e., variability or dispersion of cases, often described by the standard deviation) are helpful in assessing normality. A third characteristic, the shape of the

distribution (e.g., normal or bell-shaped, single- or multipeaked, or skewed to the left or right), is best characterized visually using histograms, box plots, and stem-and-leaf plots. Although it is important to examine, individually, the distribution of each relevant variable, it is often necessary in multivariate analyses to evaluate the pattern that exists between two or more variables. Scatterplots are a useful technique to display the shape, direction, and strength of relationships between variables.

### *Examining Atypical Data Points*

In addition to normality, data should always be preemptively examined for influential data points. Labeling observations as outside the normal range of data can be complicated because decisions exist in the context of relationships among variables and intended purpose of the data. For example, outlying $X$ values are never problematic in ANOVA designs with equal cell sizes, but they introduce significant problems in regression analyses and unbalanced ANOVA designs. However, discrepant $Y$ values are nearly always problematic. Visual detection of unusual observations is facilitated by box plots, partial regression leverage plots, partial residual plots, and influence-enhanced scatterplots. Examination of scatterplots and histograms of residual values often indicates the influence of discrepant values on the overall model fit, and whether the data point is extreme on $Y$ (outlier) or $X$ (high-leverage data point). In normally distributed data, statistical tests (e.g., $z$-score method, Leverage statistics, or Cook's $D$) can also be used to detect discrepant observations.

Some ways of handling atypical data points in normally distributed data include the use of trimmed means, scale estimators, or confidence intervals. Removal of influential observations should be guided by the research question and impact on analysis and conclusions. Sensitivity analyses can guide decisions about whether these values influence results.

## Evaluating Homoscedasticity

In addition to examining data for normality and the presence of influential data points, graphical and statistical methods are also used to evaluate homoscedasticity. These methods are often conducted as part of the analysis.

### *Regression Analyses*

In regression analysis, examination of the residual values is particularly helpful in evaluating homoscedasticity violations. The goal of regression analysis is that the model being tested will (ideally) account for all of the variation in $Y$. Variation in the residual values suggests that the regression model has somehow been misspecified, and graphical displays of residuals are informative in detecting these problems. In fact, the techniques for examining residuals are similar to those used with the original data to assess normality and the presence of atypical data points.

Scatterplots are a useful and basic graphical method to determine homoscedasticity violations. A specific type of scatterplot, known as a residual plot, plots residual $Y$ values along the vertical axis and observed or predicted $Y$ values along the horizontal ($X$) axis. If a constant spread in the residuals is observed across all values of $X$, homoscedasticity exists. Plots depicting heterosced sticity commonly show the following two patterns: (1) Residual values increase as values of $X$ increase (i.e., a right-opening megaphone pattern) or (2) residuals are highest for middle values of $X$ and decrease as $X$ becomes smaller or larger (i.e., a curvilinear relationship). Researchers often superimpose Lowess lines (i.e., lines that trace the overall trend of the data) at the mean, as well as 1 standard deviation above and below the mean of residuals, so that patterns of homoscedasticity can be more easily recognized.

### *Univariate and Multivariate Analyses of Variance*

In contexts where one or more of the independent variables is categorical (e.g., ANOVA, $t$ tests, and MANOVA), several statistical tests are often used to evaluate homoscedasticity.

In the ANOVA context, homogeneity of variance violations can be evaluated using the *FMax* and Levene's test. The *FMax* test is computed by dividing the largest variance by the smallest variance within each group. If the *FMax* exceeds the critical value found in the $F$-value table, heteroscedasticty might exist. Some researchers suggest an

*FMax* of 3.0 or more indicates a violation assumption. However, conservative estimates ($p < .025$) are suggested when evaluating *F* ratios because the *FMax* test is highly sensitive to issues of non-normality. Therefore, it is often difficult to determine whether significant values are caused by heterogeneity of variance or normality violations of the underlying population. The Levene's test is another statistical test that assumes equal variance across levels of the independent variable. If the *p* value obtained from the Levene's test is less than .05, it can be assumed that differences between variances in the population exist. Compared with the *FMax* test, the Levene's test has no required normality assumption.

In the case of MANOVA, when more than one continuous dependent variable is being assessed, the same homogeneity of variance assumption applies. Because there are multiple dependent variables, a second assumption exists that the intercorrelations among these dependent measures (i.e., covariances) are the same across different cells or groups of the design. *Box's M* test for equality of variance–covariance matrices is used to test this assumption. A statistically significant ($p < .05$) *Box's M* test indicates heteroscedasticity. However, results should be interpreted cautiously because the *Box's M* test is highly sensitive to departures from normality.

## Remediation for Violations of Homoscedasticity

### *Data Transformations*

Because homoscedasticity violations typically result from normality violations or the presence of influential data points, it is most beneficial to address these violations first. Data transformations are mathematical procedures that are used to modify variables that violate the statistical assumption of homoscedasticity. Two types of data transformations exist: linear and nonlinear transformations. Linear transformations, produced by adding, subtracting, multiplying, dividing, or a combination of these functions a constant value to the variable under consideration, preserve the relative distances of data points and the shape of the distribution. Conversely, nonlinear transformations use logs, roots, powers, and exponentials that change relative distances between data points and, therefore, influence the shape of distributions. Nonlinear transformations might be useful in multivariate analyses to normalize distributions and address homoscedasticity violations. Typically, transformations performed on *X* values more accurately address normality violations than transformations on *Y*.

Before transforming the data, it is important to determine both the extent to which the variable(s) under consideration violate the assumptions of homoscedasticity and normality and whether atypical data points influence distributions and the analysis. Examination of residual diagnostics and plots, stem-and-leaf plots, and boxplots are helpful to discern patterns of skewness, non-normality, and heteroscedasticity. In cases where a small number of influential observations is producing heteroscedasticity, removal of these few cases might be more appropriate than a variable transformation.

### *Tukey's Ladder of Power Transformations*

Tukey's ladder of power transformations ("Bulging Rule") is one of the most commonly used and simple data transformation tools. Moving up on the ladder (i.e., applying exponential functions) reduces negative skew and pulls in low outliers. Roots and logs characterize descending functions on Tukey's ladder and address problems with positive skew and high atypical data points. Choice of transformation strategy should be based on the severity of assumption violations. For instance, square root functions are often suggested to correct a moderate violation and inverse square root functions are examples of transformations that address more severe violations.

### *Advantages and Disadvantages*

There are advantages and disadvantages to conducting data transformations. First, they can remediate homoscedasticity problems and improve accuracy in a multivariate analysis. However, interpretability of results is often challenged because transformed variables are quite different from the original data values. In addition, transformations to variables where the scale range is less than 10 are often minimally effective. After performing

data transformations, it is advisable to check the resulting model because remedies in one aspect of the regression model (e.g., homoscedasticity) might lead to other model fit problems (e.g., nonlinearity).

### *Method of Weighted Least Squares*

Another remedial procedure commonly used to address heteroscedasticity in regression analysis is the method of weighted least squares (WLS). According to this method, each case is assigned a weight based on the variance of residuals around the regression line. For example, high weights are applied to data points that show a low variance of residuals around the regression line. Generally, ordinary least-squares (OLS) regression is often preferable, however, to WLS regression except in cases of large sample size or serious violations of constant variance.

*Kristen Fay*

***See also*** Homogeneity of Variance; Multivariate Normal Distribution; Normal Distribution; Normality Assumption; Normalizing Data; Parametric Statistics; Residual Plot; Variance

### Further Readings

Berry, W. D. (1993). *Understanding regression assumptions*. Newbury Park, CA: Sage.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Games, P. A., Winkler, H. B., & Probert, D. A. (1972). Robust tests for homogeneity of variance. *Educational and Psychological Measurement, 32,* 887–909.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). *Multivariate data analysis* (5th ed). Upper Saddle River, NJ: Prentice Hall.

Hartwig, F., & Dearing, B. E. (1979). *Exploratory data analysis*. Beverly Hills, CA: Sage.

Judd, C. M., McClelland, G. H., & Culhane, S. E. (1995). Data analysis: Continuing issues in the everyday analysis of psychological data. *Annual Review of Psychology, 46,* 433–465.

Meyers, L. G., Gamst, G., & Guarino, A. J. (2006). *Applied multivariate research: Design and interpretation*. London: Sage.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

# Honestly Significant Difference (HSD) Test

When an analysis of variance (ANOVA) gives a significant result, this indicates that at least one group differs from the other groups. Yet, the omnibus test does not inform on the pattern of differences between the means. To analyze the pattern of difference between means, the ANOVA is often followed by specific comparisons, and the most commonly used involves comparing two means (the so-called pairwise comparisons).

An easy and frequently used pairwise comparison technique was developed by John Tukey under the name of the *honestly significant difference* (HSD) test. The main idea of the HSD is to compute the honestly significant difference (i.e., the HSD) between two means using a statistical distribution defined by Student and called the $q$ distribution. This distribution gives the *exact* sampling distribution of the largest difference between a set of means originating from the same population. All pairwise differences are evaluated using the same sampling distribution used for the largest difference. This makes the HSD approach quite conservative.

## Notations

The data to be analyzed comprise $A$ groups; a given group is denoted $a$. The number of observations of the $a$. The group is denoted $S_a$. If all groups have the same size, it is denoted $S$. The total number of observations is denoted $N$. The mean of Group $a$ is denoted $M_{a+}$. Obtained from a preliminary ANOVA, the error source (i.e., within group) is denoted $S(A)$, the effect (i.e., between group) is denoted $A$. The mean square of error is denoted $MS_{S(A)}$, and the mean square of effect is denoted $MS_A$.

## Least Significant Difference

The rationale behind the HSD technique comes from the observation that, when the null hypothesis is true, the value of the $q$ statistics evaluating the difference between Groups $a$ and $a'$ is equal to

$$q = \frac{M_{a+} - M_{a'+}}{\sqrt{\frac{1}{2} MS_{S(A)} (\frac{1}{S_a} + \frac{1}{S_{a'}})}}, \quad (1)$$

and follows a Studentized range $q$ distribution with a range of $A$ and $N - A$ degrees of freedom. The ratio $t$ would therefore be declared significant at a given $\alpha$ level if the value of $q$ is larger than the critical value for the $\alpha$ level obtained from the $q$ distribution and denoted $q_{A,\alpha}$ where $\nu = N - A$ is the number of degrees of freedom of the error, and $A$ is the range (i.e., the number of groups). This value can be obtained from a table of the Studentized range distribution. Rewriting Equation 1 shows that a difference between the means of Group $a$ and $a'$ will be significant if

$$|M_{a+} - M_{a'+}| > \text{HSD} = q_{A,\alpha} \sqrt{\frac{1}{2} MS_{S(A)} (\frac{1}{S_a} + \frac{1}{S_{a'}})}. \quad (2)$$

When there is an equal number of observation per group, Equation 2 can be simplified as

$$\text{HSD} = q_{A,\alpha} \sqrt{\frac{MS_{S(A)}}{S}}. \quad (3)$$

To evaluate the difference between the means of Groups $a$ and $a'$, the absolute value of the difference between the means is taken and compared with the value of HSD. If

**Table 1** Results for a Fictitious Replication of Loftus & Palmer (1974) in Miles per Hour

| | *Contact* | *Hit* | *Bump* | *Collide* | *Smash* |
|---|---|---|---|---|---|
| | 21 | 23 | 35 | 44 | 39 |
| | 20 | 30 | 35 | 40 | 44 |
| | 26 | 34 | 52 | 33 | 51 |
| | 46 | 51 | 29 | 45 | 47 |
| | 35 | 20 | 54 | 45 | 50 |
| | 13 | 38 | 32 | 30 | 45 |
| | 41 | 34 | 30 | 46 | 39 |
| | 30 | 44 | 42 | 34 | 51 |
| | 42 | 41 | 50 | 49 | 39 |
| | 26 | 35 | 21 | 44 | 55 |
| *M.+* | 30 | 35 | 38 | 41 | 46 |

*Source:* Adapted from Loftus & Palmer (1974).

**Table 2** ANOVA Results for the Replication of Loftus & Palmer (1974)

| *Source* | *df* | *SS* | *MS* | *F* | *Pr(F)* |
|---|---|---|---|---|---|
| Between: *A* | 4 | 1,460.00 | 365.00 | 4.56 | .0036 |
| Error: *S(A)* | 45 | 3,600.00 | 80.00 | | |
| Total | 49 | 5,060.00 | | | |

*Source:* Adapted from Loftus & Palmer (1974).

$$|M_{a+} - M_{a'+}| \geq HSD, \quad (4)$$

then the comparison is declared significant at the chosen $\alpha$ level (usually .05 or .01). Then this procedure is repeated for all $\frac{A(A-1)}{2}$ comparisons.

Note that HSD has less power than almost all other post hoc comparison methods (e.g., Fisher's

**Table 3** HSD

| | | *Experimental Group* | | | | |
|---|---|---|---|---|---|---|
| | | $M_{1.+}$ Contact 30 | $M_{2.+}$ Hit 1 35 | $M_{3.+}$ Bump 38 | $M_{4.+}$ Collide 41 | $M_{5.+}$ Smash 46 |
| $M_{1.+} = 30$ | Contact | 000 | 5.00 ns | 8.00 ns | 11.00 ns | 16.00** |
| $M_{2.+} = 35$ | Hit | | 0.00 | 3.00 ns | 6.00 ns | 11.00 ns |
| $M_{3.+} = 38$ | Bump | | | 0.00 | 3.00 ns | 8.00 ns |
| $M_{4.+} = 41$ | Collide | | | | 0.00 | 5.00 ns |
| $M_{5.+} = 46$ | Smash | | | | | 0.00 |

*Source:* Difference between means and significance of pairwise comparisions from the (fictitious) replication of Loftus & Palmer (1974).

*Notes:* Differences larger than 11.37 are significant at the $\alpha = .05$ level and are indicated with *, and differences larger than 13.86 are significant at the $\alpha = .01$ level and are indicated with **.

LSD or Newmann–Keuls) except the Scheffé approach and the Bonferonni method because the $\alpha$ level for each difference between means is set at the same level as the largest difference.

### Example

In a series of experiments on eyewitness testimony, Elizabeth Loftus wanted to show that the wording of a question influenced witnesses' reports. She showed participants a film of a car accident and then asked them a series of questions. Among the questions was one of five versions of a critical question asking about the speed the vehicles were traveling:

1. How fast were the cars going when they *hit* each other?
2. How fast were the cars going when they *smashed into* each other?
3. How fast were the cars going when they *collided with* each other?
4. How fast were the cars going when they *bumped* each other?
5. How fast were the cars going when they *contacted* each other?

The data from a fictitious replication of Loftus's experiment are shown in Table 1. We have $A = 4$ groups and $S = 10$ participants *per* group.

The ANOVA found an effect of the verb used on participants' responses. The ANOVA table is shown in Table 2.

For an $\alpha$ level of .05, the value of $q_{.05,A}$ is equal to 4.02 and the HSD for these data is computed as

$$HSD = q_{\alpha,A}\sqrt{\frac{MS_{S(A)}}{S}} = 4.02 \times \sqrt{8} = 11.37. \quad (5)$$

The value of $q_{.01,A} = 4.90$, and a similar computation will show that, for these data, the HSD for an $\alpha$ level of .01 is equal to $HSD = 4.90 \times \sqrt{8} = 13.86$.

For example, the difference between $M_{\text{contact+}}$ and $M_{\text{hit+}}$ is declared nonsignificant because

$$|M_{\text{contact+}} - M_{\text{hit+}}| = |30 - 35| = 5 < 11.37. \quad (6)$$

The differences and significance of all pairwise comparisons are shown in Table 3.

*Hervé Abdi and Lynne J. Williams*

***See also*** Analysis of Variance (ANOVA); Bonferroni Procedure; Fisher's Least Significant Difference Test; Multiple Comparison Tests; Newman–Keuls Test and Tukey Test; Pairwise Comparisons; Post Hoc Comparisons; Scheffé Test

#### Further Readings

Abdi, H., Edelman, B., Valentin, D., & Dowling, W. J. (2009). *Experimental design and analysis for psychology.* Oxford, UK: Oxford University Press.

Hayter, A. J. (1986). The maximum familywise error rate of Fisher's least significant difference test. *Journal of the American Statistical Association, 81, 1001–1004.*

Seaman, M. A., Levin, J. R., & Serlin, R. C. (1991). New developments in pairwise multiple comparisons: Some powerful and practicable procedures. *Psychological Bulletin, 110, 577–586.*

## Hypothesis

A hypothesis is a provisional idea whose merit requires further evaluation. In research, a hypothesis must be stated in operational terms to allow its soundness to be tested.

The term *hypothesis* derives from the Greek (ὑπόθεσις), which means "to put under" or "to suppose." A scientific hypothesis is not the same as a scientific theory, even though the words *hypothesis* and *theory* are often used synonymously in common and informal usage. A theory might start as a hypothesis, but as it is subjected to scrutiny, it develops from a single testable idea to a complex framework that although perhaps imperfect has withstood the scrutiny of many research studies.

This entry discusses the role of hypotheses in research design, the types of hypotheses, and writing hypothesis.

### Hypothesis in Research Design

Two major elements in the design of research are the researcher's hypotheses and the variables to test them. The hypotheses are usually extensions

of existing theory and past research, and they motivate the design of the study. The variables represent the embodiment of the hypotheses in terms of what the researcher can manipulate and observe.

A hypothesis is sometimes described as an educated guess. However, this statement is also questioned to be a good description of hypothesis. For example, many people might agree with the hypothesis that *an ice cube will melt in less than 30 minutes if put on a plate and placed on a table*. However, after doing quite a bit of research, one might learn about how temperature and air pressure can change the state of water and restate the hypothesis as *an ice cube will melt in less than 30 minutes in a room at sea level with a temperature of 20°C or 68°F*. If one does further research and gains more information, the hypothesis might become *an ice cube made with tap water will melt in less than 30 minutes in a room at sea level with a temperature of 20°C or 68°F*. This example shows that a hypothesis is not really just an educated guess. It is a tentative explanation for an observation, phenomenon, or scientific problem that can be tested by further investigation. In other words, a hypothesis is a tentative statement about the expected relationship between two or more variables. The hypothesis is tentative because its accuracy will be tested empirically.

## Types of Hypotheses

### *Null Hypothesis*

In statistics, there are two types of hypotheses: null hypothesis ($H_0$) and alternative/research/maintained hypothesis ($H_a$). A null hypothesis ($H_0$) is a falsifiable proposition, which is assumed to be true until it is shown to be false. In other words, the null hypothesis is presumed true until statistical evidence, in the form of a hypothesis test, indicates it is highly unlikely. When the researcher has a certain degree of confidence, usually 95% to 99%, that the data do not support the null hypothesis, the null hypothesis will be rejected. Otherwise, the researcher will fail to reject the null hypothesis.

In scientific and medical applications, the null hypothesis plays a major role in testing the significance of differences in treatment and control groups. Setting up the null hypothesis is an essential step in testing statistical significance. After formulating a null hypothesis, one can establish the probability of observing the obtained data.

### *Alternative Hypothesis*

The alternative hypothesis and the null hypothesis are the two rival hypotheses whose likelihoods are compared by a statistical hypothesis test. For example, an alternative hypothesis can be a statement that the means, variance, and so on, of the samples being tested are not equal. It describes the possibility that the observed difference or effect is true. The classic approach to decide whether the alternative hypothesis will be favored is to calculate the probability that the observed effect will occur if the null hypothesis is true. If the value of this probability ($p$ value) is sufficiently small, then the null hypothesis will be rejected in favor of the alternative hypothesis. If not, then the null hypothesis will not be rejected.

### *Examples of Null Hypothesis and Alternative Hypothesis*

If a two-tailed alternative hypothesis is *that application of Educational Program A will influence students' mathematics achievements* ($H_a : \mu_{\text{Program A}} \neq \mu_{\text{control}}$), the null hypothesis is *that application of Program A will have no effect on students' mathematics achievements* ($H_0$: $\mu_{\text{Program A}} = \mu_{\text{control}}$). If a one-tailed alternative hypothesis is *that application of Program A will increase students' mathematics achievements* ($H_a : \mu_{\text{Program A}} > \mu_{\text{control}}$), the null hypothesis remains *that use of Program A will have no effect on students' mathematics achievements* ($H_0 : \mu_{\text{Program A}} = \mu_{\text{control}}$). It is not merely the opposite of the alternative hypothesis—that is, it is not *that the application of Program A will not lead to increased mathematics achievements in students*. However, this does remain the true null hypothesis.

## Hypothesis Writing

What makes a good hypothesis? Answers to the following three questions can help guide

hypothesis writing: (1) Is the hypothesis based on the review of the existing literature? (2) Does the hypothesis include the independent and dependent variables? (3) Can this hypothesis be tested in the experiment? For a good hypothesis, the answer to every question should be "Yes."

Some statisticians argue that the null hypothesis cannot be as general as indicated earlier. They believe the null hypothesis must be exact and free of vagueness and ambiguity. According to this view, the null hypothesis must be numerically exact—it must state that a particular quantity or difference is equal to a particular number.

Some other statisticians believe that it is desirable to state direction as a part of null hypothesis or as part of a null hypothesis/alternative hypothesis pair. If the direction is omitted, then it will be quite confusing to interpret the conclusion if the null hypothesis is not rejected. Therefore, they think it is better to include the direction of the effect if the test is one-sided, for the sake of overcoming this ambiguity.

*Jie Chen, Neal Kingston, Gail Tiemann, and Fei Gu*

***See also*** Directional Hypothesis; Nondirectional Hypotheses; Null Hypothesis; Research Hypothesis; "Sequential Tests of Statistical Hypotheses"

**Further Readings**

Agresti, A., & Finlay, B. (2008). *Statistical methods for the social sciences* (4th ed.). San Francisco, CA: Dellen.

Nolan, S. A., & Heinzen, T. E. (2008). *Statistics for the behavioral sciences* (3rd ed.). New York: Worth.

Shavelson, R. J. (1998). *Statistical reasoning for the behavioral sciences* (3rd ed.). Needham Heights, MA: Allyn & Bacon.

Slavin, R. (2007). *Educational research in an age of accountability*. Upper Saddle River, NJ: Pearson.

# I

## Inclusion Criteria

Inclusion criteria are a set of predefined characteristics used to identify subjects who will be included in a research study. Inclusion criteria, along with exclusion criteria, make up the selection or eligibility criteria used to rule in or out the target population for a research study. Inclusion criteria should respond to the scientific objective of the study and are critical to accomplish it. Proper selection of inclusion criteria will optimize the external and internal validity of the study, improve its feasibility, lower its costs, and minimize ethical concerns; specifically, good selection criteria will ensure the homogeneity of the sample population, reduce confounding, and increase the likelihood of finding a true association between exposure/intervention and outcomes. In prospective studies (cohort and clinical trials), they also will determine the feasibility of follow-up and attrition of participants. Stringent inclusion criteria might reduce the generalizability of the study findings to the target population, hinder recruitment and sampling of study subjects, and eliminate a characteristic that might be of critical theoretical and methodological importance.

Each additional inclusion criterion implies a different sample population and will add restrictions to the design, creating increasingly controlled conditions, as opposed to everyday conditions closer to real life, thus influencing the utility and applicability of study findings. Inclusion criteria must be selected carefully based on a review of the literature, in-depth knowledge of the theoretical framework, and the feasibility and logistic applicability of the criteria. Often, research protocol amendments that change the inclusion criteria will result in two different sample populations that might require separate data analyses with a justification for drawing composite inferences.

The selection and application of inclusion criteria also will have important consequences on the assurance of ethical principles; for example, including subjects based on race, gender, age, or clinical characteristics also might imply an uneven distribution of benefits and harms, threats to the autonomy of subjects, and lack of respect. Not including women, children, or the elderly in the study might have important ethical implications and diminish the compliance of the study with research guidelines such as those of the National Institutes of Health in the United States for inclusion of women, children, and ethnic minorities in research studies.

Use of standardized inclusion criteria is necessary to accomplish consistency of findings across similar studies on a research topic. Common inclusion criteria refer to demographic, socioeconomic, health and clinical characteristics, and outcomes of study subjects. Meeting these criteria requires screening eligible subjects using valid and reliable measurements in the form of standardized exposure and outcome measurements to ensure that

subjects who are said to meet the inclusion criteria really have them (sensitivity) and those who are said not to have them really do not have them (specificity). Such measurements also should be consistent and repeatable every time they are obtained (reliability). Good validity and reliability of inclusion criteria will help minimize random error, selection bias, misclassification of exposures and outcomes, and confounding. Inclusion criteria might be difficult to ascertain; for example, an inclusion criterion stating that "subjects with type II diabetes mellitus and no other conditions will be included" will require, in addition to clinical ascertainment of type II diabetes mellitus, evidence that subjects do not have cardiovascular disease, hypertension, cancer, and so on, which will be costly, unfeasible, and unlikely to rule out completely. A similar problem develops when using as inclusion criterion "subjects who are in good health" because a completely clean bill of health is difficult to ascertain. Choosing inclusion criteria with high validity and reliability will likely improve the likelihood of finding an association, if there is one, between the exposures or interventions and the outcomes; it also will decrease the required sample size. For example, inclusion criteria such as tumor markers that are known to be prognostic factors of a given type of cancer will be correlated more strongly with cancer than unspecific biomarkers or clinical criteria. Inclusion criteria that identify demographic, temporal, or geographic characteristics will have scientific and practical advantages and disadvantages; restricting subjects to male gender or adults might increase the homogeneity of the sample, thus helping to control confounding. Inclusion criteria that include selection of subjects during a certain period of time might overlook important secular trends in the phenomenon under study, but not establishing a feasible period of time might make conducting of the study unfeasible. Geographic inclusion criteria that establish selecting a population from a hospital also might select a biased sample that will preclude the generalizability of the findings, although it might be the only alternative to conducting the study. In studies of rheumatoid arthritis, including patients with at least 12 tender or swollen joints will make difficult the recruitment of a sufficient number of patients and will likely decrease the generalizability of study results to the target population.

The selection of inclusion criteria should be guided by ethical and methodological issues; for example, in a clinical trial to treat iron deficiency anemia among reproductive-age women, including women to assess an iron supplement therapy would not be ethical if women with life-threatening or very low levels of anemia are included in a nontreatment arm of a clinical trial for follow-up with an intervention that is less than the standard of care. Medication washout might be established as an inclusion criterion to prevent interference of a therapeutic drug on the treatment under study. In observational prospective studies, including subjects with a disease to assess more terminal clinical endpoints without providing therapy also would be unethical, even if the population had no access to medical care before the study.

In observational studies, inclusion criteria are used to control for confounding, in the form of specification or restriction, and matching. Specification or restriction is a way of controlling confounding; potential confounder variables are eliminated from the study sample, thus removing any imbalances between the comparison groups. Matching is another strategy to control confounding; matching variables are defined by inclusion criteria that will homogenize imbalances between comparison groups, thus removing confounding. The disadvantage is that variables eliminated by restriction or balanced by matching will not be amenable to assessment as potential risk factors for the outcome at hand. This also will limit generalizability, hinder recruitment, and require more time and resources for sampling. In studies of screening tests, inclusion criteria should ensure the selection of the whole spectrum of disease severity and clinical forms. Including limited degrees of disease severity or clinical forms will likely result in biased favorable or unfavorable assessment of screening tests.

Sets of recommended inclusion criteria have been established to enhance methodological rigor and comparability between studies; for example, the American College of Chest Physicians and Society of Critical Care Medicine developed inclusion criteria for clinical trials of sepsis; new criteria rely on markers of organ dysfunction rather than on blood culture positivity or clinical signs and symptoms. Recently, the Scoliosis Research Society in the United States has proposed new

standardized inclusion criteria for brace studies in the treatment of adolescent idiopathic scoliosis. Also, the International Campaign for Cures of Spinal Cord Injury Paralysis has introduced inclusion and exclusion criteria for the conduct of clinical trials for spinal cord injury. Standardized inclusion criteria must be assessed continuously because it might be possible that characteristics used as inclusion criteria change over time; for example, it has been shown that the median numbers of swollen joints in patients with rheumatoid arthritis has decreased over time. Drawing inferences using old criteria that are no longer valid would no longer be relevant for the current population.

In case control studies, inclusion criteria will define the subjects with the disease and those without it; cases and controls should be representative of the diseased and nondiseased subjects in the target population. In occupational health research, it is known that selecting subjects in the work setting might result in a biased sample of subjects who are healthier and at lower risk than the population at large. Selection of controls must be independent of exposure status, and nonparticipation rates might introduce bias. Matching by selected variables will remove the confounding effects of those variables on the association under study; obviously, those variables will not be assessed as predictors of the outcome. Matching also might introduce selection bias, complicate recruitment of subjects, and limit inference to the target population. Additional specific inclusion criteria will be needed for nested case control and case cohort designs and two-stage or multistage sampling. Common controls are population controls, neighborhood controls, hospital or registry controls, friends, relatives, deceased controls, and proxy respondents. Control selection usually requires that controls remain disease free for a given time interval, the exclusion of controls who become incident cases, and the exclusion of controls who develop diseases other than the one studied, but that might be related to the exposure of interest.

In cohort studies, the most important inclusion criterion is that subjects do not have the disease outcome under study. This will require ascertainment of disease-free subjects. Inclusion criteria should allow efficient accrual of study subjects, good follow-up participation rates, and minimal attrition.

Inclusion criteria for experimental studies involve different considerations than those for observational studies. In clinical trials, inclusion criteria should maximize the generalizability of findings to the target population by allowing the recruitment of a sufficient number of individuals with expected outcomes, minimizing attrition rates, and providing a reasonable follow-up time for effects to occur.

Automatized selection and standardization of inclusion criteria for clinical trials using electronic health records has been proposed to enhance the consistency of inclusion criteria across studies.

*Eduardo Velasco*

***See also*** Bias; Confounding; Exclusion Criteria; Reliability; Sampling; Selection; Sensitivity; Specificity; Validity of Research Conclusions

**Further Readings**

Gordis, L. (2008). *Epidemiology* (4th ed.). Philadelphia: W. Saunders.

Hulley, S. B., Cummings, S. R., Browner, W. S., Grady, D., Hearst, N., & Newman, T. B. (2007). *Designing clinical research: An epidemiologic approach* (3rd ed.). Philadelphia: Wolters, Kluwer, Lippincott, Williams & Wilkins.

LoBiondo-Wood G., & Haber J. (2006). *Nursing research: Methods and critical appraisal for evidence-based practice* (6th ed.). St. Louis, MO: Mosby.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton-Mifflin.

Szklo, M., & Nieto F. J. (2007). *Epidemiology: Beyond the basics*. Sudbury, MA: Jones & Bartlett Publishers.

## Independent Variable

Independent variable is complementary to dependent variable. These two concepts are used primarily in their mathematical sense, meaning that the value of a dependent variable changes in response to that of an independent variable. In research design, independent variables are those that a researcher can manipulate, whereas dependent variables are the responses to the effects of independent variables. By purposefully manipulating

the value of an independent variable, one hopes to cause a response in the dependent variable.

As such, independent variables might carry different names in various research fields, depending on how the relationship between the independent and the dependent variable is defined. They might be called explanatory variables, controlled variables, input variables, predictor variables, factors, treatments, conditions, or other names. For instance, in regression experiments, they often are called regressors in relation to the regressand, the dependent, or the response variable.

The concept of independent variable in statistics should not be confused with the concept of independent random variable in probability theories. In the latter case, two random variables are said to be independent if and only if their joint probability is the product of their marginal probabilities for every pair of real numbers taken by the two random variables. In other words, if two random variables are truly independent, the events of one random variable have no relationship with the events of the other random variable. For instance, if a fair coin is flipped twice, a head occurring in the first flip has no association with whether the second flip is a head or a tail because the two events are independent.

Mathematically, the relationship between independent and dependent variables might be understood in this way:

$$y = f(x), \tag{1}$$

where $x$ is the independent variable (i.e., any argument to a function) and $y$ is the dependent variable (i.e., the value that the function is evaluated to). Given an input of $x$, there is a corresponding output of $y, x$ changes independently, whereas $y$ responds to any change in $x$.

Equation 1 is a deterministic model. For each input in $x$, there is one and only one response in $y$. A familiar graphic example is a straight line if there is only one independent variable of order 1 in the previous model. In statistics, however, this model is grossly inadequate. For each value of $x$, there is often a population of $y$, which follows a probability distribution. To reflect more accurately this reality, the preceding equation is revised accordingly:

$$E(y) = f(x), \tag{2}$$

where $E(y)$ is the expectation of $y$, or equivalently,

$$y = f(x) + \varepsilon, \tag{3}$$

where $\varepsilon$ is a random variable, which follows a specific probability distribution with a zero mean. This is a probabilistic model. It is composed of a deterministic part $[f(x)]$ and a random part $(\varepsilon)$. The random part is the one that accounts for the variation in $y$.

In experiments, independent variables are the design variables that are predetermined by researchers before an experiment is started. They are carefully controlled in controlled experiments or selected in observational studies (i.e., they are manipulated by the researcher according to the purpose of a study). The dependent variable is the effect to be observed and is the primary interest of the study. The value of the dependent variable varies subjecting to the variation in the independent variables and cannot be manipulated to establish an artificial relationship between the independent and dependent variables. Manipulation of the dependent variable invalidates the entire study.

Because they are controlled or preselected and are usually not the primary interest of a study, the value of independent variables is almost universally not analyzed. Instead, they are simply taken as prescribed. (This, however, does not preclude the numeric description of independent variables as can be routinely seen in scientific literature. In fact, they are often described in detail so that a published study can be evaluated properly or repeated by others.) In contrast, the value of the dependent variable is unknown before a study. The observed value of the dependent variable usually requires careful analyses and proper explanation after a study is done.

A caution note must be sounded that even though the value of independent variables can be manipulated, one should not change it in the middle of a study. Doing so drastically modifies the independent variables before and after the change, causing a loss of the internal validity of the study. Even if the value of the dependent variable does not change drastically in response to a manipulation such, the result remains invalid. Careful

selection and control of independent variables before and during a study is fundamental to both the internal and the external validity of that study.

To illustrate what constitutes a dependent variable and what is an independent variable, let us assume an agricultural experiment on the productivity of two wheat varieties that are grown under identical or similar field conditions. Productivity is measured by tons of wheat grains produced per season per hectare. In this experiment, variety would be the independent variable and productivity the dependent variable. The qualifier, "identical or similar field conditions," implies other extraneous (or nuisance) factors (i.e., covariates) that must be controlled, or taken account of, in order for the results to be valid. These other factors might be the soil fertility, the fertilizer type and amount, irrigation regime, and so on. Failure to control or account for these factors could invalidate the experiment. This is an example of controlled experiments. Similar examples of controlled experiments might be the temperature effect on the hardness of a type of steel and the speed effect on the crash result of automobiles in safety tests.

Consider also an epidemiological study on the relationship between physical inactivity and obesity in young children: The parameter(s) that measures physical inactivity, such as the hours spent on watching television and playing video games, and the means of transportation to and from daycares/schools is the independent variable. These are chosen by the researcher based on his or her preliminary research or on other reports in literature on the same subject prior to the study. The parameter(s) that measure obesity, such as the body mass index, is (are) the dependent variable. To control for confounding, the researcher needs to consider, other than the main independent variables, any covariate that might influence the dependent variable. An example might be the social economical status of the parents and the diet of the families.

Independent variables are predetermined factors that one controls and/or manipulates in a designed experiment or an observational study. They are design variables that are chosen to incite a response of a dependent variable. Independent variables are not the primary interest of the experiment; the dependent variable is.

*Shihe Fan*

***See also*** Bivariate Regression; Covariate; Dependent Variable

**Further Readings**

Hockling, R. R. (2003). *Methods and applications of linear models: Regression and the analysis of variance*. Hoboken, NJ: Wiley.

Kuehl, R. O. (1994). *Statistical principles of research design and analysis*. Belmont, CA: Wadsworth.

Montgomery, D. C. (2001). *Design and analysis of experiments* (5th ed.). Toronto, Ontario, Canada: Wiley.

Ramsey, F. L., & Schafer, D. W. (2002). *The statistical sleuth: A course in methods of data analysis* (2nd ed.). Pacific Grove, CA: Duxbury.

Wacherly, D. D., Mendenhall , W., III, & Scheaffer, R. L. (2002). *Mathematical statistics with applications* (6th ed.). Pacific Grove, CA: Duxbury.

Zolman, J. F. (1993). *Biostatistics: Experimental design and statistical inference*. New York: Oxford University Press.

## Inference: Deductive and Inductive

Reasoning is the process of making inferences—of drawing conclusions. Students of reasoning make a variety of distinctions regarding how inferences are made and conclusions are drawn. Among the oldest and most durable of them is the distinction between *deductive* and *inductive* reasoning, which contrasts conclusions that are logically implicit in the claims from which they are drawn with those that go beyond what is given.

Deduction involves reasoning from the general to the particular:

*All mammals nurse their young.*

*Whales are mammals.*

*Therefore whales nurse their young.*

Induction involves reasoning from the particular to the general:

*All the crows I have seen are black.*

*Being black must be a distinguishing feature of crows.*

## Implication Versus Inference

Fundamental to an understanding of deductive reasoning is a distinction between *implication* and *inference*. Implication is a logical relationship; inference is a cognitive act. Statements imply; people infer. *A* implies *B* if it is impossible for *A* to be true if *B* is false. People are said to make an inference when they justify one claim (*conclusion*) by appeal to others (*premises*). Either implications exist or they do not, independently of whether inferences are made that relate to them. Inferences either are made or are not made; they can be valid or invalid, but they are inferences in either case. Failure to keep the distinction in mind can cause confusion. People are sometimes said to imply when they make statements with the intention that their hearers will see the implications of those statements and make the corresponding inferences, but to be precise in the use of language one would have to say not that people imply but that they make statements that imply.

## Aristotle and the Syllogism

The preeminent name in the history of deductive reasoning is that of Aristotle, whose codification of implicative relationships provided the foundation for the work of many generations of logicians and epistemologists. Aristotle analyzed the various ways in which valid inferences can be drawn with the structure referred to as a *categorical syllogism,* which is a form of argument involving three assertions, the third of which (the conclusion) follows from the first two (the major and minor premises). A syllogism is said to be valid if, and only if, the conclusion follows from (is implied by) the premises. Aristotle identified many valid forms and related them to each other in terms of certain properties such as *figure* and *mood*. Figure relates to the positions of the middle term—the term that is common to both premises—and mood to the types of premises involved. An explanation of the system Aristotle used to classify syllogistic forms can be found in any introductory text on first-order predicate logic.

That deductive reasoning makes explicit only knowledge already contained implicitly in the premises from which the deductions are made prompts the question: Of what practical use is deduction? One answer is that what is implicit in premises is not always apparent until it has been made explicit. It is not the case that deductive reasoning never produces surprises for those who use it. A mathematical theorem is a conclusion of a deductive argument. No theorem contains information that is not implicit in the axioms of the system from which it was derived. For many theorems, the original derivation (proof) is a cognitively demanding and time-consuming process, but once the theorem has been derived, it is available for use without further ado. If it were necessary to derive each theorem from basic principles every time it was used, mathematics would be a much less productive enterprise.

Another answer is that a conclusion is easier to retain in memory than the premises from which it was deduced and, therefore, more readily accessible for future reference. Retrieval of the conclusion from memory generally requires less cognitive effort than would the retrieval of the supporting premises and derivation of their implications.

## Validity, Plausibility, and Truth

The validity of a logical argument does not guarantee the truth of the argument's conclusion. If at least one premise of the argument is false, the conclusion also might be false, although it is not necessarily so. However, if all the argument's premises are true and its form is valid, the conclusion must be true. A false conclusion cannot follow from true premises. This is a powerful fact. Truth is consistent: Whatever a collection of true premises implies must be true. It follows that if one knows the conclusion to a valid argument to be false, one can be sure that at least one of the argument's premises is false.

Induction leads to conclusions that state more than is contained implicitly in the claims on which those conclusions are based. When, on the basis of noticing that all the members that one has seen of a certain class of things has a particular attribute ("all the crows I have seen are black"), one concludes that *all,* or *nearly all,* the members of that class (including those not seen) have that attribute, one is generalizing—going beyond the data in hand—which is one form of induction.

There are many formal constructs and tools to facilitate construction and assessment of deductive

arguments. These tools include syllogistic forms, calculi of classes and propositions, Boolean algebra, and a variety of diagrammatic aids to analysis such as truth tables, Euler diagrams, and Venn diagrams. Induction does not lend itself so readily to formalization; indeed (except in the case of mathematical induction, which is really a misnamed form of deduction) inductive reasoning is almost synonymous with informal reasoning. It has to do with weighing evidence, judging plausibility, and arriving at uncertain conclusions or beliefs that one can hold with varying degrees of confidence. Deductive arguments can be determined to be valid or invalid. The most one can say about an inductive argument is that it is more or less convincing.

Logic is often used to connote deductive reasoning only; however, it can be sufficiently broadly defined to encompass both deductive and inductive reasoning. Sometimes a distinction is made between formal and informal logic, to connote deductive and inductive reasoning, respectively.

Philosophers and logicians have found it much easier to deal with deductive than with inductive reasoning, and as a consequence, much more has been written about the former than about the latter, but the importance of induction is clearly recognized. Induction has been called the despair of the philosopher, but no one questions the necessity of using it.

Many distinctions similar to that between deductive and inductive reasoning have been made. Mention of two of them will suffice to illustrate the point. American philosopher/mathematician/logician Charles Sanders Peirce drew a contrast between a *demonstrative* argument, in which the conclusion is true whenever the premises are true, and a *probabilistic* argument, in which the conclusion is usually true whenever the premises are true. Hungarian/American mathematician George Pólya distinguished between *demonstrative* reasoning and *plausible* reasoning, demonstrative reasoning being the kind of reasoning by which mathematical knowledge is secured, and plausible reasoning that which we use to support conjectures. Ironically, although Pólya equated demonstrative reasoning with mathematics and described all reasoning outside of mathematics as plausible reasoning, he wrote extensively, especially in his 1954 two-volume *Mathematics and Plausible Reasoning,* about the role of guessing and conjecturing in mathematics.

## The Interplay of Deduction and Induction

Any nontrivial cognitive problem is almost certain to require the use of both deductive and inductive inferencing, and one might find it difficult to decide, in many instances, where the dividing line is between the two. In science, for example, the interplay between deductive and inductive reasoning is continual. Observations of natural phenomena prompt generalizations that constitute the stuff of hypotheses, models, and theories. Theories provide the basis for the deduction of predictions regarding what should be observed under specified conditions. Observations are made under the conditions specified, and the predictions are either corroborated or falsified. If falsification is the result, the theories from which the predictions were deduced must be modified and this requires inductive reasoning—guesswork and more hypothesizing. The modified theories provide the basis for deducing new predictions. And the cycle goes on.

In mathematics, a similar process occurs. A suggestive pattern is observed and the mathematician induces a conjecture, which, in some cases, becomes a theorem—which is to say it is proved by rigorous deduction from a specified set of axioms. Mathematics textbooks spend a lot of time on the proofs of theorems, emphasizing the deductive side of mathematics. What might be less apparent, but no less crucial to the doing of mathematics, is the considerable guesswork and induction that goes into the identification of conjectures that are worth exploring and the construction of proofs that will be accepted as such by other mathematicians.

Deduction and induction are essential also to meet the challenges of everyday life, and we all make extensive use of both, which is not to claim that we always use them wisely and well. The psychological research literature documents numerous ways in which human reasoning often leads to conclusions that cannot be justified either logically or empirically. Nevertheless, that the type of reasoning that is required to solve structured problems for the purposes of experimentation in the psychological laboratory does not always adequately represent the reasoning that is required to

deal with the problems that present themselves in real life has been noted by many investigators, and it is reflected in contrasts that are drawn between pure (or theoretical) and practical thinking, between academic and practical intelligence, between formal and everyday reasoning, and between other distinctions of a similar nature.

### The Study of Inferencing

The study of deductive reasoning is easier than the study of inductive reasoning because there are widely recognized rules for determining whether a deductive argument is valid, whereas there are not correspondingly widely recognized rules for determining whether an inductive argument is sound. Perhaps, as a consequence, deductive reasoning has received more attention from researchers than has inductive reasoning.

Several paradigms for investigating deduction have been used extensively by students of cognition. None is more prominent than the "selection task" invented by British psychologist Peter Wason in the 1960s. In its simplest form, a person is shown four cards, laid out so that only one side of each card is visible, and is told that each card has either a vowel or a consonant on one side and either an even number or an odd number on the other side. The visible sides of the cards show a vowel, a consonant, an even number, and an odd number. The task is to specify which card or cards must be turned over to determine the truth or falsity of the claim *If there is a vowel on one side, there is an even number on the other.* The correct answer, according to conditional logic, is the card showing a vowel and the one showing an odd number. The original finding was that only a small minority of people given this task perform it correctly; the most common selections are either the card showing a vowel and the one showing an even number, or only the one showing a vowel. The finding has been replicated many times and with many variations of the original task. Several interpretations of the result have been proposed. That the task remains a focus of research more than 60 years after its invention is a testament to the ingenuity of its inventor and to the difficulty of determining the nature of human reasoning.

*Raymond S. Nickerson*

***See also*** Experimental Design; Falsifiability; Hypothesis; Margin of Error; Nonexperimental Designs; Pre-Experimental Designs; Quasi-Experimental Designs

#### Further Readings

Cohen, L. J. (1970). *The implications of induction.* London: Methuen.

Galotti, K. M. (1989). Approaches to studying formal and everyday reasoning. *Psychological Bulletin, 105,* 331–351.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction: Processes of inference, learning, and discovery.* Cambridge: MIT Press.

Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction.* Hillside, NJ: Lawrence Erlbaum.

Pólya, G. (1954). *Mathematics and plausible reasoning, Vol. 1: Induction and analogy in mathematics, Vol. 2: Patterns of plausible inference.* Princeton, NJ: Princeton University Press.

Rips, L. J. (1994). *The psychology of proof: Deductive reasoning in human thinking.* Cambridge: MIT Press.

## Influence Statistics

Influence statistics measure the effects of individual data points or groups of data points on a statistical analysis. The effect of individual data points on an analysis can be profound, and so the detection of unusual or aberrant data points is an important part of nearly every analysis. Influence statistics typically focus on a particular aspect of a model fit or data analysis and attempt to quantify how the model changes with respect to that aspect when a particular data point or group of data points is included in the analysis. In the context of linear regression, where the ideas were first popularized in the 1970s, a variety of influence measures have been proposed to assess the impact of particular data points.

The popularity of influence statistics soared in the 1970s because of the proliferation of fast and relatively cheap computing, a phenomenon that allowed the easy examination of the effects of individual data points on an analysis for even relatively large data sets. Seminal works by R. Dennis Cook; David A. Belsley, Edwin Kuh, and Roy E. Welsch; and R. Dennis Cook and Sanford Weisberg led the way for an avalanche of new

techniques for assessing influence. Along with these new techniques came an array of names for them: DFFITS, DFBETAS, COVRATIO, Cook's $D$, and leverage, to name but a few of the more prominent examples. Each measure was designed to assess the influence of a data point on a particular aspect of the model fit: DFFITS on the fitted values from the model, DFBETAS on each individual regression coefficient, COVRATIO on the estimated residual standard error, and so on. Each measure can be readily computed using widely available statistical packages, and their use as part of an exploratory analysis of data is very common.

This entry first discusses types of influence statistics. Then we describe the calculation and limitations of influence statistics. Finally, we conclude with an example.

## Types

Influence measures are typically categorized by the aspect of the model to which they are targeted. Some commonly used influence statistics in the context of linear regression models are discussed and summarized next. Analogs are also available for generalized linear models and for other more complex models, although these are not described in this entry.

Influence with respect to fitted values of a model can be assessed using a measure called DFFITS, a scaled difference between the fitted values for the models fit with and without each individual respective data point:

$$\text{DFFITS}_i = \frac{\hat{Y}_i - \hat{Y}_{i,-i}}{\sqrt{MSE_{(i)}h_{ii}}},$$

where the notation in the numerator denotes fitted values for the response for models fit with and without the $i$th data point, respectively, $MSE_{(i)}$ is the mean square for error in the model fit without data point $i$, and $h_{ii}$ is the $i$th leverage; that is, the $i$th diagonal element of the hat matrix, $H = X(X^TX)^{-1}X^T$. Although $\text{DFFITS}_i$ resembles a $t$ statistic, it does not have a $t$ distribution, and the size of $\text{DFFITS}_i$ is judged relative to a cutoff proposed by Belsley, Kuh, and Welsch. A point is regarded as potentially influential with respect to fitted values if $|\text{DFFITS}_i| > 2\sqrt{p/n}$, where $n$ is the sample size and $p$ is the number of estimated regression coefficients.

Influence with respect to estimated model coefficients can be measured either for individual coefficients, using a measure called DFBETAS, or through an overall measure of how individual data points affect estimated coefficients as a whole. DFBETAS is a scaled difference between estimated coefficients for models fit with and without each individual datum, respectively:

$$\text{DFBETAS}_{k,i} = \frac{\hat{\beta}_k - \hat{\beta}_{k,-i}}{\sqrt{MSE_{(i)}c_{kk}}}, \text{ for } k = 1, \ldots, p,$$

where $p$ is the number of coefficients and $c_{kk}$ is the $k$th diagonal element of the matrix $(X^TX)^{-1}$. Again, although $\text{DFBETAS}_{k,i}$ resembles a $t$ statistic, it fails to have a $t$ distribution, and its size is judged relative to a cutoff proposed by Belsley, Kuh, and Welsch whereby the $i$th point is regarded as influential with respect to the $k$th estimated coefficient if $|\text{DFBETAS}_{k,i}| > 2/\sqrt{n}$.

Cook's distance calculates an overall measure of distance between coefficients estimated using models with and without each respective data point:

$$D_i = \frac{(\hat{\beta}_k - \hat{\beta}_{k,-i})^T X^T X(\hat{\beta}_k - \hat{\beta}_{k,-i})}{pMSE}.$$

There are several rules of thumb commonly used to judge the size of Cook's distance in assessing influence, with some practitioners using relative standing among the values of the $D_i$s, whereas others prefer to use the 50% critical point of the $F_{p,n-p}$ distribution.

Influence with respect to the estimate of residual standard error in a model fit can be assessed using a quantity COVRATIO that measures the change in the estimate of error spread between models fit with and without the $i$th data point:

$$\text{COVRATIO}_i = \left(\frac{s_{(i)}}{s}\right)^{2p}\left(\frac{1}{1-h_{ii}}\right),$$

where $s_{(i)}$ is the estimate of residual standard error from a model fit without the $i$th data point. Influence with respect to residual scale is assessed if a point has a value of $\text{COVRATIO}_i$ for which $|\text{COVRATIO}_i - 1| \geq 3p/n$.

Many influence measures depend on the values of the leverages, $h_{ii}$, which are the diagonal elements of the hat matrix. The leverages are a function of the explanatory variables alone and, therefore, do not depend on the response variable at all. As such, they are not a direct measure of influence, but it is observed in a large number of situations that cases having high leverage tend to be influential. The leverages are closely related to the Mahalanobis distances of each data point's covariate values from the centroid of the covariate space, and so points with high leverage are in that sense "far" from the center of the covariate space. Because the average of the leverages is equal to $p/n$, where $p$ is the number of covariates plus 1, it is common to consider points with twice the average leverage as having the potential to be influential; that is, points with $h_{ii} > 2p/n$ would be investigated further for influence. Commonly, the use of leverage in assessing influence would occur in concert with investigation of other influence measures.

## Calculation

Although the formulas given in the preceding discussion for the various influence statistics are framed in the context of models fit with and without each individual data point in turn, the calculation of these statistics can be carried out without the requirement for multiple model fits. This computational saving is particularly important in the context of large data sets with many covariates, as each influence statistic would otherwise require $n+1$ separate model fits in its calculation. Efficient calculation is possible through the use of updating formulas. For example, the values of $s_{(i)}$, the residual standard error from a model fit without the $i$th data point, can be computed via the formula

$$(n-p-1)s^2_{(i)} = (n-p)s^2 - \frac{e_i^2}{1-h_{ii}},$$

where $s$ is the residual standard error fit using the entire data set and $e_i$ is the model errors from the model fit to the full data set. Similarly,

$$\text{DFFITS}_i = \frac{e_i\sqrt{h_{ii}}}{s_{(i)}(1-h_{ii})}$$

and similar expressions not requiring multiple model fits can be developed for the other influence measures considered earlier.

## Limitations

Each influence statistic discussed so far is an example of a *single-case deletion* statistic, based on comparing models fit on data sets differing by only one data point. In many cases, however, more than one data point in a data set exerts influence, either individually or jointly. Two problems that can develop in the assessment of multiple influence are *masking* and *swamping*. Masking occurs when an influential point is not detected because of the presence of another, usually adjacent, influential point. In such a case, single-case deletion influence statistics fail because only one of the two potentially influential points is deleted, respectively, when computing the influence statistic, still leaving the other data point to influence the model fit. Swamping occurs when "good" data points are identified as influential because of the presence of other, usually remote, influential data points that influence the model away from the "good" data point. It is difficult to overcome the potential problems of masking and swamping for several reasons: First, in high-dimensional data, visualization is often difficult, making it very hard to "see" which observations are "good" and which are not; second, it is almost never the case that the exact number of influential points is known *a priori*, and points might exert influence either individually or jointly in groups of unknown size; and third, multiple-case deletion methods, although simple in conception, remain difficult to implement in practice because of the computational burden associated with assessing model fits for very large numbers of subsets of the original data.

## Examples

A simple example concludes this entry. A "good" data set with 20 data points was constructed, to which was added, first, a single obvious influential point, and then a second, adjacent influential point. The first panel of Figure 1 depicts the original data, and the second and third panels show the augmented data. An initial analysis of the "good" data reveals no points suspected of being

**Table 1** Values of Influence Statistics for Example Data

| *Two influential points added (right panel of figure)* | | | | | | | |
|---|---|---|---|---|---|---|---|
| *Point* | *DFFITS* | *DFBETAS (0)* | *DFBETAS (1)* | *Cook's* D | *COVRATIO* | *Leverage* | |
| 4 | –1 | –0.993 | 0.455 | 0.283 | 0.335 | 0.0572 | Swamped |
| 21 | –0.406 | 0.13 | –0.389 | 0.086 | 2.462 | 0.5562 | Masked |
| 22 | –0.148 | 0.042 | –0.14 | 0.012 | 1.973 | 0.4399 | Masked |
| Cutoff | 0.603 | 0.426 | 0.426 | 0.718 | (0.73, 1.27) | 0.18 | |
| *One influential point added (middle panel of figure)* | | | | | | | |
| *Point* | *DFFITS* | *DFBETAS (0)* | *DFBETAS (1)* | *Cook's* D | *COVRATIO* | *Leverage* | |
| 4 | –0.984 | –0.966 | 0.419 | 0.273 | 0.338 | 0.0582 | Swamped |
| 21 | –126 | 49.84 | –126.02 | 1100 | 2.549 | 0.9925 | Highly influential |
| Cutoff | 0.617 | 0.436 | 0.436 | 0.719 | (0.71, 1.29) | 0.18 | |

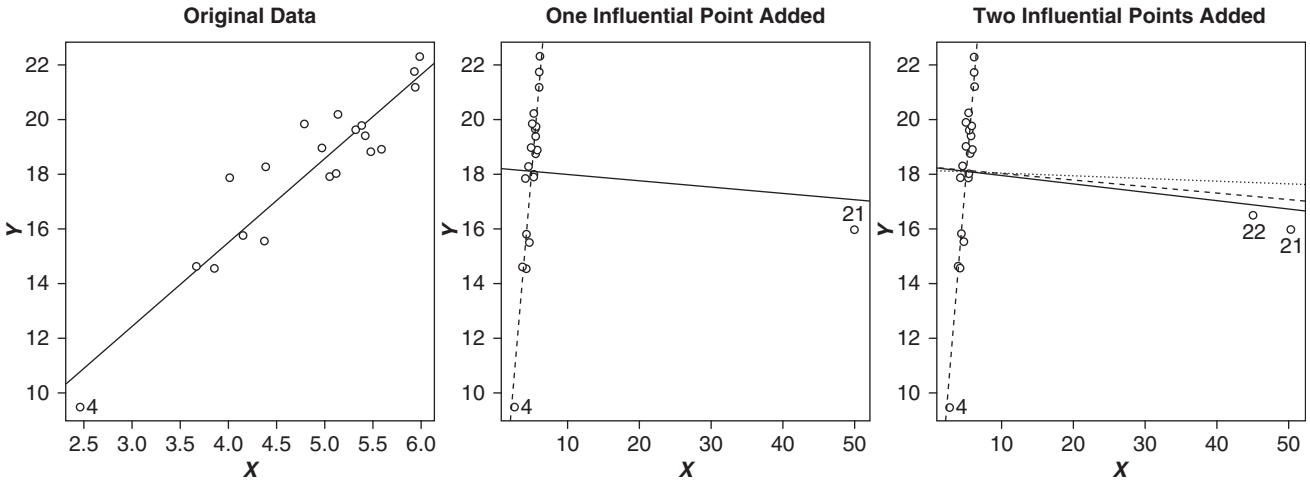



**Figure 1** Influence Statistics Example Data

influential. When the first influential point is inserted in the data, its impact on the model is extreme (see the middle plot), and the influence statistics clearly point to this point as being influential. When the second extreme point is added (see the rightmost plot), its presence obscures the influence of the initially added point (point 22 masks point 21, and vice versa), and the pair of added points causes a known "good" point, labeled 4, to be considered influential (the pair (21,22) swamps point 4). In the plots, the fitted model using all data points is marked using a solid line, whereas the fitted model using only the "good" data points is marked using a dashed line. In the rightmost plot, the dotted line reflects the fitted model using all data points except point 22, whereas the dot-dash line reflects the fitted model using all data points except point 21. Of course, in this simple example, the effects of the data points marked in the plot are clearly visible—the simple two-dimensional case usually affords such an easy visualization. In higher dimensions, such visualization is typically not possible, and so the values of influence statistics become more useful as tools for identifying unusual or influential data points.

Table 1 shows the values of the various influence statistics for the example depicted in the figure. Values of DFFITS, DFBETAS, Cook's *D*, COVRATIO, and leverage are given for the situations depicted in the middle and right panels of the figure. The values of the influence statistics for the case of the single added influential point show

how effectively the influence statistics betray the added influential point—their values are extremely high across all statistics. The situation is very different, however, when a second, adjacent influential point is added. In that case, the two added points mask each other, and at the same time, they swamp a known "good" point. The dotted line and the dot-dash line in the rightmost panel of Figure 1 clearly show how the masking occurs—the fitted line is barely changed when either of the points 21 or 22 is individually removed from the data set. These points exert little individual influence, but their joint influence is extreme.

*Michael A. Martin and Steven Roberts*

***See also*** Data Cleaning; Data Mining; Outlier; SPSS

**Further Readings**

Atkinson, A. C., & Riani, M. (2000). *Robust diagnostic regression analysis.* New York: Springer-Verlag.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity.* New York: Wiley.

Chatterjee, S., & Hadi, A. S. (1986). Influential observations, high leverage points, and outliers in linear regression. *Statistical Science, 1,* 379–393.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics, 19,* 15–18.

Cook, R. D. (1979). Influential observations in linear regression. *Journal of the American Statistical Association, 74,* 169–174.

Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression.* New York: Chapman & Hall.

# Influential Data Points

Influential data points are observations that exert an unusually large effect on the results of regression analysis. Influential data might be classified as outliers, as leverage points, or as both. An *outlier* is an anomalous response value, whereas a *leverage point* has atypical values of one or more of the predictors. It is important to note that not all outliers are influential.

Identification and appropriate treatment of influential observations are crucial in obtaining a valid descriptive or predictive linear model. A single, highly influential data point might dominate the outcome of an analysis with hundreds of observations: It might spell the difference between rejection and failure to reject a null hypothesis or might drastically change estimates of regression coefficients. Assessing influence can reveal data that are improperly measured or recorded, and it might be the first clue that certain observations were taken under unusual circumstances. This entry discusses the identification and treatment of influential data points.

## Identifying Influential Data Points

A variety of straightforward approaches is available to identify influential data points on the basis of their leverage, outlying response values, or individual effect on regression coefficients.

### *Graphical Methods*

In the case of simple linear regression ($p = 2$), a contingency plot of the response versus predictor values might disclose influential observations, which will fall well outside the general two-dimensional trend of the data. Observations with high leverage as a result of the joint effects of multiple explanatory variables, however, are difficult to reveal by graphical means. Although simple graphing is effective in identifying extreme outliers and nonsensical values, and is valuable as an initial screen, the eyeball might not correctly discern less obvious influential points, especially when the data are sparse (i.e., small $n$).

### *Leverage*

Observations whose influence is derived from explanatory values are known as leverage points. The *leverage* of the $i$th observation is defined as $h_i = \mathbf{x}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}'_i$, where $\mathrm{x}_i$ is the $i$th row of the $n \times p$ design matrix $\mathbf{X}$ for $p$ predictors and sample size $n$. Larger values of $h_i$, where $0 \leq h_i \leq 1$, are indicative of greater leverage. For reasonably large data sets ($n - p > 50$), a value of $h_i$ greater than $2p/n$ is a standard criterion for classification as a leverage point, where $\sum_{i=1}^{n} h_i = p$ and thus the mean of $h_i = p / n$.

### *Standardized Residuals*

An objective test for outliers is available in the form of standardized residuals. The Studentized deleted residuals, $e_i^* = \frac{y_i - \hat{y}_i}{s_{(i)}\sqrt{1-h_i}}$, where $s^2_{(i)}$ is the mean square estimate of the residual variance $\sigma^2$ with the $i$th observation removed, have a Student's $t$ distribution with $n - p - 1$ degrees of freedom ($df$) under the assumption of normally distributed errors. An equivalent expression might be constructed in terms of $y_i(i)$, which is the fitted value for observation $i$ when the latter is not included in estimating regression parameters: $e_i^* = \frac{1-h_i}{h_i}\ [\hat{y}_i - \hat{y}_i(i)]$. As a rule of thumb, an observation might be declared an outlier if $|e_i^*| > 3$. As mentioned, however, classification as an outlier does not necessarily imply large influence.

### *Estimates of Influence*

Several additional measures assess influence on the basis of effect on the model fit and estimated regression parameters. The standardized change in fit, $\text{DFFITS}_i = \frac{\sqrt{h_i}(y_i - \hat{y}_i)}{(1-h_i)s_{(i)}}$, provides a standardized measure of effect of the $i$th observation on its *fitted* (*predicted*) *value*. It represents the change, in units of standard errors (*SE*), in the fitted value brought about by omission of the $i$th point in fitting the linear model. DFFITS and the Studentized residual are closely related: $\text{DFFITS}_i = e_i^* \sqrt{h_i/(1-h_i)}$. The criteria for large effect are typically $|\text{DFFITS}| > 2\sqrt{p/n}$ for large data sets or $|\text{DFFITS}| > 1$ for small data sets. This measure might be useful where prediction is the most important goal of an analysis.

More generally, it is of interest to examine effect on estimated regression coefficients. Influence on individual parameters might be assessed through a standardized measure of change: $\text{DFBETAS}_{ij} = \frac{\hat{\beta}_j - \hat{\beta}_j(i)}{s_{(i)}\sqrt{(\mathbf{X}'\mathbf{X})^{-1}_{jj}}}$, where $\hat{\beta}_j$ and $\hat{\beta}_j(i)$ are the least-squares estimates of the $j$th coefficient with and without the $i$th data point, respectively. DFBETAS, like DFFITS, measures effect in terms of the estimated *SE*. A reasonable criterion for a high level of influence is $|\text{DFBETAS}_{ij}| > 2/\sqrt{n}$.

A composite score for influence on all coefficient estimates is available in Cook's distance, $D_i = (\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}\ \ (\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})/ps^2 = \frac{h_i(y_i - \hat{y}_i)^2}{(1-h_i)^2 ps^2}$, where $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}(i)$ are the $p \times 1$ vectors of parameter estimates with and without observation $i$, respectively. Cook's distance scales the distance between $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}(i)$ such that under the standard assumptions of linear regression, a value greater than the median of an $F$ distribution with $p$ and $n - p$ $df$ is generally considered to be highly influential.

### *Multiple Influential Points*

The measures described previously are geared toward finding single influential data points. In a few cases, they will fail to detect influential observations because two or more similarly anomalous points might conceal one another's effect. Such situations are quite unusual, however. Similar tests have been developed that are generalized to detect multiple influential points simultaneously.

## Treatment of Influential Data Points

Although influential data points should be carefully examined, they should not be removed from the analysis unless they are unequivocally proven to be erroneous. Leverage points that nevertheless have small effect might be beneficial, as they tend to enhance the precision of coefficient estimates. Observations with large effect on estimation can be acknowledged, and results both in the presence and absence of the influential observations can be reported. It is possible to downweight influential data points without omitting them completely, for example, through weighted linear regression or by Winsorization. *Winsorization* reduces the influence of outliers without completely removing observations by adjusting response values more centrally. This approach is appropriate, for example, in genetic segregation and linkage analysis, in which partially centralizing extreme values scales down their influence without changing inference on the underlying genotype.

Alternatively, and especially when several valid but influential observations are found, one might consider *robust regression;* this approach is relatively insensitive to even a substantial percentage of outliers. A large number of atypical or outlying

values might indicate an overall inappropriateness of the linear model for the data. For example, it might be necessary to transform (normalize) one or more variables.

*Robert P. Igo, Jr.*

***See also*** Bivariate Regression; Data Cleaning; Exploratory Data Analysis; Graphical Display of Data; Influence Statistics; Residual Plot; Robust; Winsorize

**Further Readings**

Belsley, D. A., Kuh, E., & Welsch, R. E. (2004). *Regression diagnostics: Identifying influential data and sources of collinearity.* Hoboken, NJ: Wiley.

Cook, R. D. (1977). Detection of influential observations in linear regression. *Technometrics, 19,* 15–18.

Cook, R. D., & Weisberg, S. (1994). *An introduction to regression graphics.* Hoboken, NJ: Wiley.

Fox, J. (1997). *Applied regression analysis, linear models and related methods.* Thousand Oaks, CA: Sage.

# Informed Consent

Protecting human participants in research is extremely important, and part of that process is informed consent. Informed consent is an ongoing communication process between research participants and the investigator to ensure participants' comfort. Informed consent allows potential research participants to volunteer their participation freely, without threat or undue coaching. The potential participant is also provided with information an individual would want to know before participating, so an educated decision can be made whether or not to participate. Generally, the purpose of informed consent is to protect each participant's welfare, ensure the participants are voluntary and informed, and promote positive feelings before and after completing a study.

This entry begins with a brief history and then describes the necessary components of informed consent and some additional considerations. Next, the entry discusses the methods involved in obtaining informed consent, including special cases. The entry concludes with a discussion of situations in which exceptions to informed consent process might be made.

## History

Today, protecting human research participants through informed consent is common practice. That was not always the case. Unethical and harmful studies conducted in the past led to the creation of a regulatory board and current ethical principles that are in place to protect the rights of human participants.

The Nazis conducted a great deal of inhuman research during World War II. As a result, in 1947 the Nuremberg Military Tribunal created the Nuremberg Code, which protected human participants in medical experiments. The code required researchers to obtain voluntary consent and minimize harm in experiments that would provide more benefits to the participants than foreseen risks. In 1954, the National Institutes of Health (NIH) established an ethics committee that adopted a policy that required all human participants to provide voluntary informed consent. Furthermore, the Department of Health, Education, and Welfare issued regulations in 1974 that called for protection of human research participants. The department would not support any research that was not first reviewed and approved by a committee, which is now known as the Institutional Review Board (IRB). The IRB would be responsible for determining the degree of risk and whether the benefits outweighed any risk to the participants. It was noted in the regulations that informed consent must be obtained.

The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research was created in 1974, and this organization was charged to identify basic ethical principles within research with human subjects and make recommendations to improve the policies in place. As a result, the Belmont Report was created, which identifies and defines the three basic principles for research, and it is still in use today. The basic principles of the Belmont report include respect for persons, beneficence, and justice. Beneficence is the process of maximizing good outcomes from science, humanity, and the research participants, while avoiding or minimizing unnecessary risk, harm, or wrong. Respect encompasses the overall protection of an individual's autonomy, through courtesy and respect for everyone, including individuals who are not autonomous (children,

mentally handicapped, etc.). Justice ensures reasonable, nonexploitative, and carefully considered procedures through fair administration.

As a result of the Belmont Report, six norms were determined for conducting research: valid research designs, competence of researcher, identification of consequences, selection of subjects, voluntary informed consent, and compensation for injury. Each norm coexists with the others to ensure participants' safety and should be followed by researchers when formulating and implementing a research project.

Several revisions were made to the ethics code as time progressed. More recently, the Common Rule was created and applied. The Common Rule established the following three main protective factors: review of research by an IRB, institutional assurances of compliance, and informed consent of participants. The Common Rule is also still used today.

## Required Components

Generally, three conditions must be met for informed consent to be considered valid—the participants must understand the information presented, the consent must be given voluntary, and the participant must be competent to give consent. More specifically, federal law requires eight components be included in the consent statement.

First, an explanation of the purpose of the research, the expected duration of the subject's participation, and a description of the procedure must be included. Details of the methods are not required and are actually discouraged to allow clearer comprehension on the part of the participant. Jargon, legal terminology, and irrelevant information should not be included. If deception is necessary, then participants must be informed that the details of the study cannot be explained prior to the study, and that they will be given a full explanation of the study upon completion. Additional information on the use of deception is discussed later.

Second, any description of foreseeable risk or discomfort should be explained. Risk implies that harm, loss, or damages might occur. This can include mere inconvenience or physical, psychological, social, economic, and legal risks.

Third, a description of any benefits to the subjects or others that are expected should be explained. Benefits include scientific knowledge; personally relevant benefits for the participants (i.e., food, money, and medical/mental health services); insight, training, learning, role modeling, empowerment, and future opportunities; psychosocial benefits (i.e., altruism, favorable attention, and increased self-esteem); kinship benefits (i.e., closeness to people or reduction of alienation); and community benefits (i.e., policies and public documentation).

Fourth, descriptions of alternatives to participate must be provided to potential participants. This provides additional resources to people who are being recruited.

The fifth requirement is a description of how confidentiality or anonymity will be ensured and its limits. Anonymity can be ensured in several ways. Examples include using numbers or code names instead of the names of the participants. The specifics of the study will likely determine how confidentiality will be ensured.

Sixth, if the research will have more than minimal risk, law requires a statement of whether compensation for injury will be provided. If compensation will be provided, a description of how should be included.

The seventh requirement is to provide the contact information of the individual(s) the participants can contact if they have any questions or if there was the event of harm.

Eighth, a statement will be made that participation is voluntary and that if one chooses not to participate, there will be no penalty or loss. In addition, it must be explained that if one chooses to participate, then leaving the study at any time is acceptable and there would be no penalty.

Last, the participants should receive a copy of the informed consent to keep.

## Other Considerations

There are several elements that can be added to an informed consent form to make it more effective, although these items are not required. Examples include the following: any circumstances that might warrant termination of a participant regardless of consent, additional costs the participants might experience, the procedure if a participant

decided to leave the study and its consequences, and developments of the study.

Overall, an effective consent statement should be jargon free, easy to understand, and written in a friendly, simple manner. A lengthy description of the methods is not necessary, and any irrelevant information should not be included. As discussed previously, each legal requirement should be included as well.

In addition to the content and the style in which the informed consent is written, the manner in which the material is presented can increase or decrease participation. Establishing good rapport is very important and might require specific attention, as presenting the informed consent might become mundane if repeated a great deal. Using a friendly greeting and tone throughout the process of reading the informed consent is important. Using body language that displays openness will be helpful as well. A lack of congruence between what is verbalized and displayed through body language might lead potential participants to feel uncomfortable. Furthermore, using an appropriate amount of eye contact will help create a friendly atmosphere as well. Too little or too much eye contact could potentially be offensive to certain individuals. Presenting a willingness to answer all concerns and questions is important as well. Overall, potential participants will better trust researchers who present themselves in a friendly, caring manner and who create a warm atmosphere.

## Methods of Obtaining Informed Consent

There are several methods in which consent can be obtained. Largely, consent is acquired through written (signed) consent. Oral and behavioral consent are other options that are used less commonly.

In most cases, the IRB will require a signed consent form. A signed consent form provides proof that consent was indeed obtained. In riskier studies, having a witness sign as well can provide extra assurance.

The actual written consent form can take two forms—one that contains each required element outlined previously or a short written consent document. If the full version is presented, the form can be read to or by the potential participants. The short form entails documenting that the required criteria of an informed consent were read orally to the participant or the participants' legally authorized representative. The IRB must also approve a written summary of what will be said orally to the potential participants. Only the short form will be signed by the participant. In addition, a witness must sign the short form and the summary of what was presented. A copy of the short form and the written summary should be provided to the participant. Whichever way the material is presented, the individual should be provided adequate time to consider the material before signing.

Behavioral consent occurs when the consent form is waived or exempt. These situations are discussed later.

## Special Cases

Several cases require special considerations in addition to the required components of informed consent. These special cases include minors, individuals with disabilities, language barriers, third parties, studies using the Internet for collection, and the use of deception in research.

To protect children and adolescent research participants, safeguards are put into place. Children might be socially, cognitively, or psychologically immature, and therefore, cannot provide informed consent. In 1983, the Department of Health and Human Services adopted a federal regulation governing behavioral research on persons under the age of 18. The regulations that were put into place include several components. First, an IRB approval must be obtained. Next, the documented permission of one parent or guardian and the assent of the child must be obtained. Assent is the child's affirmative agreement to participate in the study. A lack of objection is not enough to assent. The standard for assent is the child's ability to understand the purpose and what will occur if one chooses to participate. In the case of riskier research, both parents' permission must be obtained. Furthermore, the research must involve no greater risk than the child normally encounters, unless the risk is justified by anticipated benefits to the child.

Adapting the assent process with young children can lead to better comprehension. Minimizing the level of difficulty by using simple language is effective to describe the study. After the

presentation of the information, the comprehension of the child should be assessed. Repeating the material or presenting it in a story or video format can be effective as well.

Another group that requires special consideration is individuals with disabilities. Assessing mental stability and illness as well as cognitive ability is important to determine the participants' ability to make an informed decision. Moreover, considerations to the degree of impairment and level of risk are critical to ensure the requirements of informed consent are met. Often, a guardian or surrogate will be asked to provide consent for participation in a study with disabled individuals.

Cultural issues also need consideration when obtaining consent from individuals of different nationalities and ethnicities. Individuals who speak a language other than English might have difficulty understanding the material presented in the informed consent. Special considerations should be made to address this issue. For instance, an interpreter can be used or the form could be transcribed to the native language of the potential participant. Translators can reduce language barriers significantly and provide an objective presentation of the information about the study.

Protecting third parties in research, or information obtained about other people from a participant, is another special case. Although there are no guidelines currently in place, some recommendations exist. Contextual information that is obtained from participants is generally not considered private. However, when information about a third party becomes identifiable and is private, an informed consent must be obtained.

With advances in technology, many studies are being carried out through the Internet because of the efficiency and low cost to researchers. Unfortunately, ethical issues, including informed consent, are hard to manage online. Researchers agree ethical guidelines are necessary, and some recommendations have been made, but currently there is no standardized method of colleting and validating informed consent online. Concerns about obtaining consent online include being certain the participant is of legal age to consent and that the material presented was understood. Maintaining confidentiality with the use of e-mail and managing deception is difficult. Suggestions to these issues include recruiting participants via the Internet and then sending informed consent forms through the mail to obtain signatures. When the signed copy is obtained, a code can be sent back to participate in the online study. Another suggestion is using a button, where the participant has to click "I agree" after reading the informed consent. After giving consent, access to the next page would be granted.

The use of deception in research is another special concern that is extremely controversial. By definition, using deception does not meet the criteria of informed consent to provide full disclosure of information about the study to be conducted. Deception in research includes providing inaccurate information about the study, concealing information, using confederates, making false guarantees in regard to confidentiality, misrepresenting the identity of the investigator, providing false feedback to the participants, using placebos, using concealed recording devices, and failing to inform people they are part of a study. Proponents of deceptive research practice argue that deception provides useful information that could not otherwise be obtained if participants were fully informed. The American Psychological Association (APA) guidelines allow the use of deception with specific regulations. The use of deception must be justified clearly by the prospective scientific value, and other alternatives must be considered before using deception. Furthermore, debriefing must occur no later than the end of data collection to explain the use of deception in the study and all the information fully that was originally withheld.

## Exceptions

There are cases in which an IRB will approve consent procedures with elements missing or with revisions from the standard list of requirements, or in which they will waive the written consent entirely.

In general, the consent form can be altered or waived if it is documented that the research involves no more harm than minimal risk to the participants, the waiver or alteration will not adversely affect the rights and welfare of the subjects, the research could not be practically carried out without the waiver or alteration, or the subjects will be provided with additional pertinent information after participating.

Alterations and waivers can also be made if it is demonstrated that the research will be conducted or is subject to the approval of state or local government officials, and it is designed to study, evaluate, or examine (a) public benefit service programs, (b) procedures for obtaining benefits or services under these programs, (c) possible changes in or alternatives to those programs or procedures, or (d) possible changes in methods or level of payment for benefits or services under those programs. Furthermore, under the same conditions listed previously, required elements can be left out if the research could not be practically carried out without the waiver or alteration.

The IRB can also waive the requirement for the researcher to obtain a signed consent form in two cases. First, if the only record linking the participant and the research would be the consent form and the primary risk would be potential harm from a breach of confidentiality, then a signed consent form can be waived. Each participant in this case should be provided the choice of whether he or she would like documentation linking him or her to the research. The wishes of the participant should then be followed. Second, the research to be conducted presents no more than minimal risk to the participants and involves no procedures that normally require a written consent outside of a research context. In each case, after approving a waiver, the IRB could require the researcher to provide participants with a written statement explaining the research that will be conducted.

Observational studies, ethnographic studies, survey research, and secondary analysis can all waive informed consent. In observational studies, a researcher observes the interaction of a group of people as a bystander. If the participants remain anonymous, then informed consent can be waived. An ethnographic study involves the direct observation of a group through an immersed researcher. Waiving consent in ethnographic studies depends on the case and vulnerability of the participants. In conducting survey research, if the participant can hang up the phone or throw away mail, then consent is likely not needed. If a survey is conducted in person and the risk is minimal to the participant, then consent can be waived as well. Furthermore, informed consent does not have to be obtained for secondary analysis of data.

Parental permission can also be waived under two circumstances. Consent can be waived for research involving only minimal risk, given that the research will not affect the welfare of the participants adversely, and the research can be carried out practically without a waiver. For instance, children who live on the streets might not have parents who could provide consent. In addition, parental permission can also be waived if they do not properly protect the child.

The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research identified four other cases in which a waiver for parental consent could potentially occur. Research designed to study factors related to the incidence or treatment of conditions in adolescents who could legally receive treatment without parental permission is one case. The second is participants who are mature minors and the procedures involve no more risk than usual. Third, research designed to study neglected or abused children, and fourth, research involving children whose parents are not legally or functionally competent, does not require parental consent.

Child assent can also be waived if the child is deemed incapable of assenting, too young, immature, or psychologically unstable, or if obtaining assent would hinder the research possibilities. The IRB must approve the terms before dismissing assent.

There are several cases in which research is exempted from the IRB and, therefore, does not require the consent of the parent(s). One of the most common cases is if research is conducted in commonly accepted educational settings that involve normal educational practices. Examples of normal practices include research on instructional strategies and research on the effectiveness of, or the comparison among, instructional technique, curricula, or classroom management. Another common example is if research involves educational tests (i.e., cognitive, diagnostic, aptitude, or achievement), or if the information from these tests cannot be identified to the participants.

*Rhea L. Owens*

*See also* Assent; Interviewing; Observations; Participants; Recruitment

**Further Readings**

Adair, J. G., Dushenko, T. W., & Lindsey, R. C. L. (1985). Ethical regulations and their impact on research practice. *American Psychologist, 40,* 59–72.

American Psychological Association. (1992). Ethical principles of psychologist and code of conduct. *American Psychologist,* 47, 1597–1611.

Citro, C. F., Ilgen, D. R., & Marrett, C. B. (2003). *Protecting participants and facilitating social and behavioral sciences research*. Washington, DC: National Academic Press.

Collogan, L. K., & Fleischman, A. R. (2005). Adolescent research and parental permission. In E. Kodish (Ed.), *Ethics and research with children: A case-based approach*. New York: Oxford University Press.

Gross, B. (2001). Informed consent. *Annals of the American Psychotherapy Association, 4,* 24.

Jensen, P. S., Josephson, A. M., & Frey, J. (1989). Informed consent as a framework for treatment: Ethical and therapeutic considerations. *American Journal of Psychotherapy, 3,* 378–386.

Michalak, E. E., & Szabo, A. (1998). Guidelines for Internet research: An update. *European Psychologist, 3,* 70–75.

Miller, C. (2003). Ethical guidelines in research. In J. C. Thomas & M. Hersen (Eds.), *Understanding research in clinical and counseling psychology.* Mahwah, NJ: Lawrence Erlbaum.

Sieber, J. E. (1992). *Planning ethically responsible research: A guide for students and internal review boards.* Newbury Park, CA: Sage.

Tymchuk, A. J. (1991). Assent processes. In B. Stanley & J. E. Sieber (Eds.), *Social research on children and adolescents: Ethical issues*. Newbury Park, CA: Sage.

# INSTRUMENTATION

Instrumentation refers to the tools or means by which investigators attempt to measure variables or items of interest in the data-collection process. It is related not only to instrument design, selection, construction, and assessment, but also the to conditions under which the designated instruments are administered—the instrument is the device used by investigators for collecting data. In addition, during the process of data collection, investigators might fail to recognize that changes in the calibration of the measuring instrument(s) can lead to biased results. Therefore, instrumentation is also a specific term with respect to a threat to internal validity in research. This entry discusses instrumentation in relation to the data-collection process, internal validity, and research designs.

## Instrumentation Pertaining to the Whole Process of Data Collection

Instrumentation is the use of, or work completed by, planned instruments. In a research effort, it is the responsibility of an investigator to describe thoroughly the instrument used to measure the dependent variable(s), outcome(s), or the effects of interventions or treatments. In addition, because research largely relies on data collection through measurement, and instruments are assigned with operational numbers to measure purported constructs, instrumentation inevitably involves the procedure of establishing instrument validity and reliability as well as minimizing measurement errors.

### *Validity and Reliability*

Validity refers to the extent to which an instrument measures what it purports to measure with investigated subjects. Based on the research necessities, investigators need to determine ways to assess instrument validity that best fits the needs and objectives for the research. In general, instrument validity consists of face validity, content validity, criterion-related validity, and construct-related validity. It is necessary to note that an instrument is simply valid for measuring a particular purpose and for a designated group. That is, an instrument can be valid for measuring a group of specific subjects but can become invalid for another. For example, a valid 4th-grade math achievement test is unlikely to be a valid math achievement test for 2nd graders. In another instance, a valid 4th-grade math achievement test is unlikely to be a valid aptitude test for 4th graders.

Reliability refers to the degree to which an instrument consistently measures whatever the instrument was designed to measure. A reliable instrument can generate consistent results. More specifically, when an instrument is applied to target subjects more than once, an investigator can expect to obtain results that are quite similar or even identical each time. Such measurement

consistency enables investigators to gain confidence in the measuring ability or dependability of the particular instrument. Approaches to reliability consist of repeated measurements on an individual (i.e., test–retest and equivalent forms), internal consistency measures (i.e., split-half, Kuder–Richardson 20, Kuder–Richardson 21, and Cronbach's alpha), and interrater and intrarater reliability. Usually, reliability is shown in the numerical form, as a coefficient. The range of reliability coefficient is from 0 (errors existed in the entire measurement) to 1 (no error in the measurement was discovered); the higher the coefficient, the better the reliability.

### *Measurement Errors*

Investigators need to attempt to minimize measurement errors whenever practical and possible for the purpose of accurately indicating the reported values collected by the instrument. Measurement errors can occur for various reasons and might result from the conditions of testing (e.g., test procedure not properly followed, testing site too warm or too cold for subjects to calmly respond to the instrument, noise distractions, or poor seating arrangements), from characteristics of the instrument itself (e.g., statements/questions not clearly stated, invalid instruments of measuring the concept in question, unreliable instruments, or statements/questions too long), from test subjects themselves (e.g., socially desirable responses provided by subjects, bogus answers provided by subjects, or updated or correct information not possessed by subjects), or combinations of these listed errors. Pamela L. Alreck and Robert B. Settle refer to the measurement errors described previously as instrumentation bias and error.

Concerning the validity and reliability of instrumentation, measurement errors can be both systematic and random. Systematic errors have an impact on instrument validity, whereas random errors affect instrument reliability. For example, if a group of students were given a math achievement test and the test was difficult to all examinees, then all test scores would be systematically lowered. These lowered scores indicate that the validity of the math achievement test is low for that particular student group or, in other words, the instrument does not measure what it purports to measure because the performance is low for all subjects. Measurement errors can also take place in a random fashion. In this case, for example, if a math achievement test is reliable and if a student has been projected to score 70 based on her previous performance, then investigators would expect test scores of this student to be close to the projected score of 70. After the same examination was administered on several different occasions, the scores obtained (e.g., 68, 71, and 72) might not be the exact projected score—but they are pretty close. In this case, the differences in test scores would be caused by random variation. Conversely, of course, if the test is not reliable, then considerable fluctuations in terms of test scores would not be unusual. In fact, any values or scores obtained from such an instrument would be, more or less, affected by random errors, and researchers can assume that no instrument is totally free from random errors. It is imperative to note that a valid instrument *must* have reliability. An instrument can, however, be reliable but invalid—consistently measuring the wrong thing.

Collectively, instrumentation involves the whole process of instrument development and data collection. A good and responsible research effort requires investigators to specify where, when, and under what conditions the data are obtained to provide scientific results and to facilitate similar research replications. In addition to simply indicating where, when, and under what conditions the data are obtained, the following elements are part of the instrumentation concept and should be clearly described and disclosed by investigators: how often the data are to be collected, who will collect the data, and what kinds of data-collection methods are employed. In summary, instrumentation is a term referring to the process of identifying and handling the variables that are intended to be measured in addition to describing how investigators establish the quality of the instrumentation concerning validity and reliability of the proposed measures, how to minimize measurement errors, and how to proceed in the process of data collection.

## Instrumentation as a Threat to Internal Validity

As discussed by Donald T. Campbell and Julian Stanley in *Experimental and Quasi-Experimental*

*Designs for Research,* instrumentation, which is also named instrument decay, is one of the threats to internal validity. It refers to changes in calibration of a measuring instrument or changes in persons collecting the data that can adversely generate differences in the data gathered thereby affecting the internal validity of a study. This threat can result from data-collector characteristics, data-collector bias, and the decaying effect. Accordingly, certain research designs are susceptible to this threat.

### *Data Collector Characteristics*

The results of a study can be affected by the characteristics of data collectors. When more than two data collectors are employed as observers, scorers, raters, or recorders in a research project, a variety of individual characteristics (e.g., gender, age, working experience, language usage, and ethnicity) can interject themselves into the process and thereby lead to biased results. For example, this situation might occur when the performance of a given group is rated by one data collector while the performance of another group is collected by a different person. Suppose that both groups perform the task equally well per the performance criteria. However, the score of one group is significantly higher than that of the other group. The difference in raters would be highly suspect in causing the variations in measured performance. The principal controls to this threat are to use identical data collector(s) throughout the data-collection process, to analyze data separately for every data collector, to precalibrate or make certain that every data collector is equally skilled in the data collection task, or ensure that each rater has the opportunity to collect data from each group.

### *Data Collector Bias*

It is possible that data collectors might unconsciously treat certain subjects or groups differently than the others. The data or outcome generated under such conditions would inevitably produce biased results. Data collector bias can occur regardless of how many investigators are involved in the collection effort; a single data-collection agent is subject to bias. Examples of the bias include presenting "leading" questions to the persons being interviewed, allowing some subjects to use more time than others to complete a test, or screening or editing sensitive issues or comments by those collecting the data. The primary controls to this threat are to standardize the measuring procedure and to keep data collectors "blind." Principal investigators need to provide training and standardized guidelines to make sure that data collectors are aware of the importance of measurement consistency within the process of data collection. With regard to keeping data collectors "blind," principal investigators need to keep data collectors ignorant of which method individual subjects or groups (e.g., control group vs. experimental group) are being tested or observed in a research effort.

### *Decaying Effect*

When the data generated from an instrument allow various interpretations and the process of handling those interpretations are tedious and/or difficult requiring rigorous discernment, an investigator who scores or needs to provide comments on these instruments one after another can eventually become fatigued, thereby leading to scoring differences. A change in the outcome or conclusion supported by the data supports has now been introduced by the investigator, who is an extraneous source not related to the actual collected data. A common example would be that of an instructor who attempts to grade a large number of term papers. Initially, the instructor is thorough and painstaking on his or her assessment of performance. However, after grading many papers, tiredness, fatigue, and clarity of focus gradually factor in and influence his or her judgments. The instructor then becomes more generous on scoring the second half of the term papers. The principal control to this threat is to arrange several data-collection or grading sessions to keep the scorer calm, fresh, and mentally acute while administering examinations or grading papers. By doing so, the decaying effect that leads to scoring differences can be minimized.

## Instrumentation as a Threat to Research Designs

Two quasi-experimental designs, the time-series and the separate-sample pretest–posttest designs,

and one preexperimental design, the one-group pretest–posttest group design, are vulnerable to the threat of instrumentation. The time-series design is an elaboration of a series of pretests and posttests. For reasons too numerous to include here, data collectors sometimes change their measuring instruments during the process of data collection. If this is the case, then instrumentation is introduced and any main effect of the dependent variable can be misread by investigators as the treatment effect. Instrumentation can also be a potential threat to the separate-sample pretest–posttest design. Donald T. Campbell and Julian Stanley note that differences in attitudes and experiences of a single data collector could be confounded with the variable being measured. That is, when a data collector has administered a pretest, he or she would be more experienced in the posttest and this difference might lead to variations in measurement. Finally, the instrumentation threat can be one of the obvious threats often realized in the one-group pretest–posttest group design (one of the preexperimental designs). This is a result of the six uncontrolled threats to internal validity inherent with this design (i.e., history, maturation, testing, instrumentation, regression, interaction of selection, and maturation). Therefore, with only one intervention and the pre- and posttest design, there is a greater chance of being negatively affected by data collector characteristics, data collector bias, and decaying effect, which can produce confounded results. The effect of the biases are difficult to predict, control, or identify in consideration of any effort to separate actual treatment effects with the influence of these extraneous factors.

### Final Note

In the field of engineering and medical research, the term *instrumentation* is frequently used and refers to the development and employment of accurate measurement, analysis, and control. Of course, in the fields mentioned previously, instrumentation is also associated with the design, construction, and maintenance of actual instruments or measuring devices that are not proxy measures but the actual device or tool that can be manipulated per its designed function and purpose. Comparing the devices for measurement in engineering with the social sciences, the latter is much less precise. In other words, the fields of engineering might use instrumentation to collect hard data or measurements of the real world, whereas research in the social sciences produces "soft" data that only measures perceptions of the real world. One reason that instrumentation is complicated is because many variables act independently as well as interact with each other.

*Chia-Chien Hsu and Brian A. Sandford*

***See also*** Internal Validity; Reliability; Validity of Measurement

**Further Readings**

Alreck, P. L., & Settle, R. B. (1995). *The survey research handbook: Guidelines and strategies for conducting a survey.* New York: McGraw-Hill.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Fraenkel, J. R., & Wallen, N. E. (2000). *How to design and evaluate research in education.* Boston: McGraw-Hill.

Gay, L. R. (1992). *Educational research: Competencies for analysis and application.* Upper Saddle River, NJ: Prentice Hall.

## INTERACTION

In most research contexts in the biopsychosocial sciences, researchers are interested in examining the influence of two or more predictor variables on an outcome. For example, researchers might be interested in examining the influence of stress levels and social support on anxiety among first-semester graduate students. In the current example, there are two predictor variables—stress levels and social support—and one outcome variable—anxiety. In its simplest form, a statistical interaction is present when the association between a predictor and an outcome varies significantly as a function of a second predictor. Given the current example, one might hypothesize that the association between stress and anxiety varies significantly as a function of social support. More specifically, one might hypothesize that there is no association between stress and anxiety among individuals reporting higher levels of social support while

simultaneously hypothesizing that the association between stress and anxiety is strong among individuals reporting lower levels of social support. Data consistent with these joint hypotheses would be suggestive of a significant interaction between stress and social support in predicting anxiety.

Hypothetical data consistent with this interaction are presented in Figure 1. The horizontal axis is labeled *Stress,* with higher values representing higher levels of stress. The vertical axis is labeled *Anxiety,* with higher values representing higher levels of anxiety. In figures such as these, one predictor (in this case, stress) is plotted along the horizontal axis, while the outcome (in this case, anxiety) is plotted along the vertical axis. The second predictor (in this case, social support) forms the lines in the plot. In Figure 1, the flat line is labeled *High Social Support* and represents the association between stress and anxiety for individuals reporting higher levels of social support. The other line is labeled *Low Social Support* and represents the association between stress and anxiety for individuals reporting lower levels of social support. In plots like Figure 1, as the lines depart from parallelism, a statistical interaction is suggested.

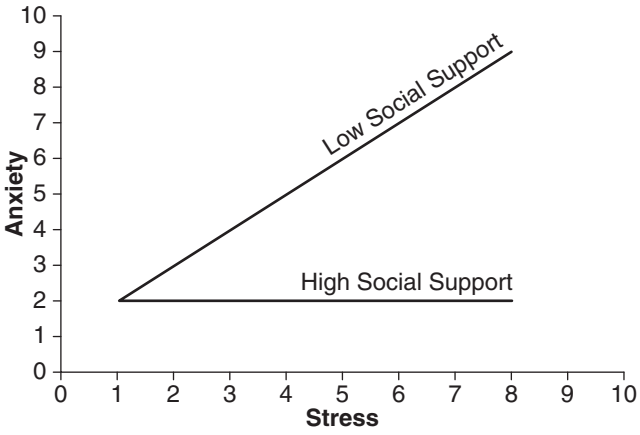



**Figure 1** The Interaction of Stress and Support in Predicting Anxiety

## Terminological Clarity

Researchers use many different terms to discuss statistical interactions. The crux issue in describing statistical interactions has to do with *dependence.* The original definition presented previously stated that a statistical interaction is present when the association between a predictor and an outcome varies significantly as a function of a second predictor. Another way of stating this is that the effects of one predictor on an outcome *depend* on the value of a second predictor. Figure 1 is a great illustration of such dependence. For individuals who report higher levels of social support, there is no association between stress and anxiety. For individuals who report lower levels of social support, there is a strong positive association between stress and anxiety. Consequently, the association between stress and anxiety *depends* on level of social support. Some other terms that researchers use to describe statistical interactions are (a) *conditional on,* (b) *contingent on,* (c) *modified by,* and/or (d) *moderated by.* Researchers might state that the effects of stress on anxiety are *contingent on* social support or that support moderates the stress–anxiety association. The term *moderator* is commonly used in various fields in the social sciences. Researchers interested in testing hypotheses involving *moderation* are interested in testing statistical interactions that involve the putative moderator and at least one other predictor. In plots like Figure 1, the moderator variable will often be used to form the lines in the plot while the remaining predictor is typically plotted along the horizontal axis.

## Statistical Models

Statistical interactions can be tested using many different analytical frameworks. For example, interactions can be tested using analysis of variance (ANOVA) models, multiple regression models, and/or logistic regression models—just to name a few. Next, we highlight two such modeling frameworks—ANOVA and multiple regression—although for simplicity, the focus is primarily on ANOVA.

### *Analysis of Variance*

The ANOVA model is often used when predictor variables can be coded as finite categorical variables (e.g., with two or three categories) and the outcome is continuous. Social science researchers who conduct laboratory-based experiments often use the ANOVA framework to test research

hypotheses. In the ANOVA framework, predictor variables are referred to as *independent variables* or *factors,* and the outcome is referred to as the *dependent variable*. The simplest ANOVA model that can be used to test a statistical interaction includes two factors—each of which has two categories (referred to as *levels* in the ANOVA vernacular). Next, an example is presented with hypothetical data that conform to this simple structure. This example assumes that 80 participants were randomly assigned to one of four conditions: (1) low stress and low support, (2) low stress and high support, (3) high stress and low support, and (4) high stress and high support. In this hypothetical study, one can assume that stress was manipulated by exposing participants to either a simple (low stress) or complex (high stress) cognitive task. One can assume also that a research confederate was used to provide either low or high levels of social support to the relevant study participant while she or he completed the cognitive task.

**Table 1** Data From a Hypothetical Laboratory-Based Study Examining the Effects of Stress and Social Support on Anxiety

| | | *Support* | | |
|---|---|---|---|---|
| | | *Low* | *High* | |
| *Stress* | *Low* | 4 | 2 | 3 |
| | *High* | 10 | 6 | 8 |
| | | 7 | 4 | |

### Main Effects and Simple Effects

A discussion of interactions in statistical texts usually involves the juxtaposition of two kinds of effects: *main effects* and *simple effects* (also referred to as *simple main effects*). Researchers examining main effects (in the absence of interactions) are interested in the unique independent effect of each of the predictors on the outcome. In these kinds of models—which are often referred to as *additive effects models*—the effect of each predictor on the outcome is constant across all levels of the remaining predictors. In sharp contrast, however, in examining models that include interactions, researchers are interested in exploring the possibility that the effects of one predictor on the outcome depend on another predictor. Next, this entry examines both main and interactive effects in the context of the hypothetical laboratory-based experiment.

In Table 1, hypothetical data from the laboratory-based study are presented. The numbers inside the body of the table are means (i.e., arithmetic averages) on the dependent variable (i.e., anxiety). Each mean is based on a distinct group of 20 participants who were exposed to a combination of the stress (e.g., low) and support (e.g., low) factors. For example, the number 10 in the body of the table is the mean anxiety score among individuals in the high-stress/low-support condition. Means are presented also in the *margins* of the tables (i.e., the underlined numbers). The *marginal means* are the means of the two relevant row or column entries. The *main effect* of a factor is examined by comparing the marginal means across the various levels of the factor. In the current example, we assume that any (nonzero) difference between the marginal means is equivalent to a main effect for the factor in question. Based on this assumption, the main effect of stress in Table 1 is significant—because there is a difference between the two stress marginal means of 3 (for the low-stress condition) and 8 (for the high-stress condition). As might be expected, on average individuals exposed to the high-stress condition reported higher levels of anxiety than did individuals exposed to the low-stress condition. Similarly, the main effect of support is also significant because there is a difference between the two support marginal means of 7 (for the low-support condition) and 4 (for the high-support condition).

In the presence of a statistical interaction, however, the researcher's attention turns away from the *main effects* of the factors and instead focuses on the *simple effects* of the factors. As described previously, in the data presented in Table 1 the main effect of support are quantified by comparing the means of individuals who received either low or high levels of social support. A close examination of Table 1 makes clear that scores contributing to the low-support mean derive from the following two different sources: (1) individuals who were exposed to a low-stress cognitive task and (2) individuals who were exposed to a high-stress cognitive task. The same is true of scores contributing to the high-support mean. In the current example, however, combining

**Table 2** Heuristic Table of Various Effects From a Hypothetical 2 × 2 Design

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| | | ***Support*** | | |
| | | *Low* | *High* | |
| ***Stress*** | *Low* | A | B | E |
| | *High* | C | D | F |
| | | G | H | |

*Note*. Cell values above (e.g., A) represent group means—each of which is presumed to be based on 20 scores. Underlined values (e.g., E) represent marginal means—each of which is the average of the relevant row or column entries (i.e., E is the average of A and B).

| *Effect* | *Contrast* |
|---|---|
| Main effect of support | H − G |
| Main effect of stress | F − E |
| Simple support at low stress | B − A |
| Simple support at high stress | D − C |
| Simple stress at low support | C − A |
| Simple stress at high support | D − B |
| Stress by support interaction | (D − C) − (B − A) or (D − B) − (C − A) |

*Note*. The minus signs in the contrast column are meant to denote subtraction.

data from individuals exposed to either lower or higher levels of stress does not seem prudent. In part, this is true because the mean anxiety score for individuals exposed to high support varies as a function of stress. In other words, when holding support constant at high levels, on average individuals exposed to the low-stress task report lower levels of anxiety (2) than do individuals exposed to the high-stress task (6). Even more important, the stress effect is more pronounced at lower levels of support because the average anxiety difference between the low-stress (4) and high-stress (10) conditions is larger. In other words, the data presented in Table 1 suggest that the effects of stress on anxiety depend on the level of support. Another way of saying this is that there is a statistical interaction between stress and support in predicting anxiety.

As noted previously, in exploring interactions, researchers focus on *simple* rather than *main* effects. Although the term was not used, two of the four simple effects yielded by the hypothetical study have already been discussed. When examining simple effects, the researcher contrasts the table means in one row or column of the table. In doing so, the researcher is examining the simple effects of one of the factors at a specific level (i.e., value) of the other factor.

Previously, when we observed that there was a difference in the mean anxiety scores for individuals who received a low level of social support under either the low-stress or high-stress conditions, we were discussing the *simple effect of stress at low levels of support*. In other words, comparing the mean of the low-support/low-stress group (4) to the mean of the low-support/high-stress group (10) is testing the simple effect of stress at low levels of support. If there is a (nonzero) difference between these two means, we will assume that the simple effect of stress at lower levels of support is statistically significant. Consequently, in the current case, this simple effect is significant (because $10 - 4 = 6$). In examining the simple effects of stress at high levels of support, we would compare the high-support/low-stress mean (2) with the high-support/high-stress mean (6). We would conclude that the simple effect of stress at high levels of support is also significant (because $6 - 2 = 4$).

The test of the interaction effect is quantified by examining whether the relevant simple effects are different from one another. If the difference between the two simple effects of stress—one at lower levels of support and the second at higher levels of support are compared—it will be found that the interaction is significant (because $4 - 6 = -2$). The fact that these two simple effects differ quantifies numerically the original conceptual definition of a statistical interaction, which stated that in its simplest form a statistical interaction is present when the association between a predictor and an outcome varies significantly as a function of a second predictor. In the current case, the association between stress and anxiety varies significantly as a function of support. At higher levels of support, the (simple) effect of stress is more muted, resulting in a mean anxiety difference of 4 between the low-stress

and high-stress conditions. At lower levels of support, however, the (simple) effect of stress is more magnified, resulting in a mean anxiety difference of 6 between the low-stress and high-stress conditions. Consequently, the answer to the question "What is the effect of stress on anxiety?" is "It depends on the level of support." (Table 2 provides a generic 2 × 2 table in which all of the various main and simple effects are explicitly quantified. The description of the Table 1 entries as well as the presentation of Table 2 should help the reader gain a better understanding of these various effects.)

### *Multiple Regression*

In 1968, Jacob Cohen shared some of his insights with the scientific community in psychology regarding the generality and flexibility of the multiple regression approach to data analysis. Within this general analytical framework, the ANOVA model exists as a special case. In the more general multiple regression model, predictor variables can take on any form. Predictors might be unordered categorical variables (e.g., gender: male or female), ordered categorical variables (e.g., symptom severity: low, moderate, or high), and/or truly continuous variables (e.g., chronological age). Similarly, interactions between and/or among predictor variables can include these various mixtures (e.g., a categorical predictor by continuous predictor interaction or a continuous predictor by continuous predictor interaction).

Many of the same concepts described in the context of ANOVA have parallels in the regression framework. For example, in examining an interaction between a categorical variable (e.g., graduate student cohort: first year or second year) and a continuous variable (e.g., graduate school–related stress) in predicting anxiety, a researcher might examine the *simple slopes* that quantify the association between stress and anxiety for each of the two graduate school cohorts. In such a model, the test of the (cohort by stress) interaction is equivalent to testing whether these simple slopes are significantly different from one another.

In 1991, Leona S. Aiken and Stephen G. West published their seminal work on testing, interpreting, and graphically displaying interaction effects in the context of multiple regression. In part, the impetus for their work derived from researchers' lack of understanding of how to specify and interpret multiple regression models properly, including tests of interactions. In some of the more commonly used statistical software programs, ANOVA models are typically easier to estimate because the actual coding of the effects included in the analysis occurs "behind the scenes." In other words, if a software user requested a full factorial ANOVA model (i.e., one including all main effects and interactions) to analyze the data from the hypothetical laboratory-based study described previously, the software would create *effect codes* to specify the stress and support predictors and would also form the product of these codes to specify the interaction predictor. The typical user, however, is probably unaware of the coding that is used to create the displayed output. In specifying a multiple regression model to analyze these data, the user would not be spared the work of coding the various effects in the analysis. More important, the user would need to understand the implications of the various coding methods for the proper interpretation of the model estimates. This is one reason why the text by Aiken and West has received so much positive attention. The text outlines—through the use of illustrative examples—various methods used to test and interpret multiple regression models including tests of interaction effects. It also includes a detailed discussion of the proper interpretation of the various conditional (i.e., simple) effects that are components of larger interactions.

## Additional Considerations

This discussion endeavored to provide a brief and nontechnical introduction to the concept of statistical interactions. To keep the discussion more accessible, equations for the various models described were not provided. Moreover, this discussion focused mostly on interactions that were relatively simple in structure (e.g., the laboratory-based example, which involved an interaction between two 2-level categorical predictors). Before concluding, however, this entry broaches some other important issues relevant to the discussion of statistical interactions.

### *Interactions Can Include Many Variables*

All the interactions described previously involve the interaction of two predictor variables. It is possible to test interactions involving three or more predictors as well (as long as the model can be properly identified and estimated). In the social sciences, however, researchers rarely test interactions involving more than three predictors. In testing more complex interactions the same core concepts apply—although they are generalized to include additional layers of complexity. For example, in a model involving a three-way interaction, seven effects comprise the full factorial model (i.e., three main effects; three 2-way interactions; and one 3-way interaction). If the three-way interaction is significant, it suggests that the *simple two-way interactions* vary significantly as a function of the third predictor.

### *When Is Testing an Interaction Appropriate?*

This discussion thus far has focused nearly exclusively on understanding simple interactions from both conceptual and statistical perspectives. When it is appropriate to test statistical interactions has not been discussed. As one might imagine, the answer to this question depends on many factors. A couple of points are worthy of mention, however. First, some models require the testing of statistical interactions in that the models assume that the predictors in question do *not* interact. For example, in the classic analysis of covariance (ANCOVA) model in which one predictor is treated as the predictor variable of primary theoretical interest and the other predictor is treated as a covariate (or statistical control variable), the model assumes that the primary predictor and covariate do not interact. As such, researchers employing such models should test the relevant (predictor by covariate) interaction as a means of assessing empirically this model assumption. Second, as is true in most empirical work in the sciences, theory should drive both the design of empirical investigations and the statistical analyses of the primary research hypotheses. Consequently, researchers can rely on the theory in a given area to help them make decisions about whether to hypothesize and test statistical interactions.

*Christian DeLucia and Brandon Bergman*

*See also* Analysis of Variance (ANOVA); Effect Coding; Factorial Design; Main Effects; Multiple Regression; Simple Main Effects

### Further Readings

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51,* 1173–1182.

Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin, 70,* 426–443.

Cohen, J. (1978). Partialed products *are* interactions; Partialed powers *are* curve components. *Psychological Bulletin, 85,* 858–866.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Keppel, G. (1991). *Design and analysis: A researcher's handbook* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

## Internal Consistency Reliability

Internal consistency reliability estimates how much total test scores would vary if slightly different items were used. Researchers usually want to measure constructs rather than particular items. Therefore, they need to know whether the items have a large influence on test scores and research conclusions.

This entry begins with a discussion of classical reliability theory. Next, formulas for estimating internal consistency are presented, along with a discussion of the importance of internal consistency. Last, common misinterpretations and the interaction of all types of reliability are examined.

### Classical Reliability Theory

To examine reliability, classical test score theory divides observed scores on a test into two components, true score and error:

$$X = T + E,$$

where $X$ = observed score, $T$ = true score, and $E$ = error score.

If Steve's true score on a math test is 73 but he gets 71 on Tuesday because he is tired, then his observed score is 71, his true score is 73, and his error score is –2. On another day, his error score might be positive, so that he scores better than he usually would.

Each type of reliability defines true score and error differently. In test–retest reliability, true score is defined as whatever is consistent from one testing *time* to the next, and error is whatever varies from one testing time to the next. In interrater reliability, true score is defined as whatever is consistent from one *rater* to the next, and error is defined as whatever varies from one rater to the next. Similarly, in internal consistency reliability, true score is defined as whatever is consistent from one *item* to the next (or one set of items to the next set of items), and error is defined as whatever varies from one item to the next (or from one set of items to the next set of items that were designed to measure the same construct). To state this another way, true score is defined as the expected value (or long-term average) of the observed scores—the expected value over many times (for test–retest reliability), many raters (for interrater reliability), or many items (for internal consistency). The true score is the average, not the truth. The error score is defined as the amount by which a particular observed score differs from the average score for that person.

Researchers assess all types of reliability using the reliability coefficient. The reliability coefficient is defined as the ratio of true score variance to observed score variance:

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2},$$

where $\rho_{XX'}$ = the reliability coefficient, $\sigma_T^2$ = the variance of true scores across participants, and $\sigma_X^2$ = the variance of observed scores across participants.

Classical test score theory assumes that true scores and errors are uncorrelated. Therefore, observed variance on the test can be decomposed into true score variance and error variance:

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2,$$

where $\sigma_E^2$ = the variance of error scores across participants.

The reliability coefficient can now be rewritten as follows:

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2}.$$

Reliability coefficients vary from 0 to 1, with higher coefficients indicating higher reliability.

This formula can be applied to each type of reliability. Thus, internal consistency reliability is the proportion of observed score variance that is caused by true differences between participants, where true differences are defined as differences that are consistent across the set of items. If the reliability coefficient is close to 1, then researchers would have obtained similar total scores if they had used different items to measure the same construct.

## Estimates of Internal Consistency

Several different formulas have been proposed to estimate internal consistency reliability. Lee Cronbach, Cyril Hoyt, and Louis Guttman independently developed the most commonly used formula, which is labeled coefficient alpha after the terminology used by Cronbach. The split-half approach is also common. In this approach, the test is divided into two halves, which are then correlated. G. F. Kuder and M. W. Richardson developed KR-20 for use with dichotomous items (i.e., true/false items or items that are marked as correct or incorrect). KR-20 is easy to calculate by hand and has traditionally been used in classroom settings. Finally, Tenko Raykov and Patrick Shrout have recently proposed measuring internal consistency reliability using structural equation modeling approaches.

## Importance of Internal Consistency

Internal consistency reliability is the easiest type of reliability to calculate. With test–retest reliability, the test must be administered twice. With interrater reliability, the test must be scored twice. But with internal consistency reliability, the test only

needs to be administered once. Because of this, internal consistency is the most commonly used type of reliability.

Internal consistency reliability is important when researchers want to ensure that they have included a sufficient number of items to capture the concept adequately. If the concept is narrow, then just a few items might be sufficient. For example, the International Personality Item Pool (IPIP) includes a 10-item measure of self-discipline that has a coefficient alpha of .85. If the concept is broader, then more items are needed. For example, the IPIP measure of conscientiousness includes 20 items and has a coefficient alpha of .88. Because conscientiousness is a broader concept than self-discipline, if the IPIP team measured conscientiousness with just 10 items, then the particular items that were included would have a substantial effect on the scores obtained, and this would be reflected in a lower internal consistency.

Second, internal consistency is important if a researcher administers different items to each participant. For example, an instructor might use a computer-administered test to assign different items randomly to each student who takes an examination. Under these circumstances, the instructor must ensure that students' course grades are mostly a result of real differences between the students, rather than which items they were assigned.

However, it is unusual to administer different items to each participant. Typically, researchers compare scores from participants who completed identical items. This is in sharp contrast with other forms of reliability. For example, participants are often tested at different times, both within a single study and across different studies. Similarly, participants across different studies are usually scored by different raters, and sometimes participants within a single study are scored by different raters. When differences between testing times or raters are confounded with differences between participants, researchers must consider the effect of this design limitation on their research conclusions. Because researchers typically only compare participants who have completed the same items, this limitation is usually not relevant to internal consistency reliability.

Thus, the internal consistency coefficient tells researchers how much total test scores would vary if different items were used. This question is theoretically important because it tells researchers whether they have covered the full breadth of the construct. But this question is usually not of practical interest, because researchers usually administer the same items to all participants.

## Common Misinterpretations

Four misinterpretations of internal consistency are common. First, researchers often assume that if internal consistency is high, then other types of reliability are high. In fact, there is no necessary mathematical relationship between the variance caused by items, the variance caused by time, and the variance caused by raters. It might be that there is little variance caused by items but considerable variance caused by time and/or raters. Because each type of reliability defines true score and error score differently, there is no way to predict one type of reliability based on another.

Second, researchers sometimes assume that high internal consistency implies unidimensionality. This misinterpretation is reinforced by numerous textbooks that state that the internal consistency coefficient indicates whether all items measure the same construct. However, Neal Schmitt showed that a test can have high internal consistency even if it measures two or more unrelated constructs. This is possible because internal consistency reliability is influenced by both the relationships between the items and the number of items. If all items are related strongly to each other, then just a few items are sufficient to obtain high internal consistency. If items have weaker relationships or if some items have strong relationships and other items are unrelated, then high internal consistency can be obtained by having more items.

Researchers often want to know whether a set of items is unidimensional, because it is easier to interpret test scores if all items measure the same construct. Imagine that a test contains 10 vocabulary items and 10 math items. Jane scores 10 by answering the 10 vocabulary items correctly; John scores 10 by answering the 10 math items correctly; and Chris scores 10 by answering half of the vocabulary and half of the math items correctly. All three individuals obtain the same score, but these identical scores do not reflect similar abilities. To avoid this problem, researchers often

want to know whether test items are unidimensional. However, as stated previously, internal consistency does not imply unidimensionality.

To determine whether items measure a unitary construct, researchers can take one of two approaches. First, they can calculate the average interitem correlation. This correlation measures how closely the items are related to each other and is the most common measure of item homogeneity. However, the average interitem correlation might disguise differences between items. Perhaps some items have strong relationships with each other and other items are unrelated. Second, researchers can determine how many constructs underlie a set of items by conducting an exploratory factor analysis. If one construct underlies the items, the researcher can determine whether some items measure that construct better than others. If two or more constructs underlie the items, then the researcher can determine which items measure each construct and create homogeneous subscales to measure each. In summary, high internal consistency does not indicate that a test is unidimensional; instead, researchers should use exploratory factor analysis to determine dimensionality.

The third misinterpretation of internal consistency is that internal consistency is important for all tests. There are two exceptions. Internal consistency is irrelevant if test items are identical and trivially easy. For example, consider a speed test of manual dexterity. For each item, participants draw three dots within a circle. Participants who complete more items within the time limit receive higher scores. When items are identical and easy, as they are in this example, J. C. Nunnally and I. H. Berstein showed that internal consistency will be very high and hence is not particularly informative. This conclusion makes sense conceptually: When items are identical, very little variance in total test scores is caused by differences in items. Researchers should instead focus on other types of reliability, such as test–retest.

Internal consistency is also irrelevant when the test is designed deliberately to contain heterogeneous content. For example, if a researcher wants to predict success in an area that relies on several different skills, then a test that assesses each of these skills might be useful. If these skills are independent of each other, the test might have low internal consistency. However, applicants who score high on the test possess all the necessary skills and might do well in the job. If few applicants score high on all items, the company might need a more detailed picture of the strengths and weaknesses of each applicant. In that case, the researcher could develop internally consistent subscales to measure each skill area, as described previously. In that case, internal consistency would be relevant to the subscales but would remain irrelevant to the total test scores.

Fourth, researchers often assume mistakenly that the formulas that are used to assess internal consistency—such as coefficient alpha—are only relevant to internal consistency. Usually these formulas are used to estimate the reliability of total (or average) scores on a set of $k$ items, but these formulas can also be used to estimate the reliability of total scores from a set of $k$ times or $k$ raters—or any composite score. For example, if a researcher is interested in examining stable differences in emotion, participants could record their mood each day for a month. The researcher could average the mood scores across the 30 days for each participant. Coefficient alpha can be used to estimate how much of the observed differences between participants are caused by differences between days and how much is caused by stable differences between the participants. Alternatively, researchers could use coefficient alpha to examine raters. Job applicants could be rated by each manager in a company, and the average ratings could be calculated for each applicant. The researcher could use coefficient alpha to estimate the proportion of variance caused by true differences between the applicants—as opposed to the particular set of managers who provided ratings. Thus, coefficient alpha (and the other formulas discussed previously) can be used to estimate the reliability of any score that is calculated as the total or average of parallel measurements—whether those parallel measurements are obtained from different items, times, or raters.

## Going Beyond Classical Test Score Theory

In classical test score theory, each source of variance is considered separately. Internal consistency reliability estimates the effect of test items. Test–retest reliability estimates the effect of time. Interrater reliability estimates the effect of rater. To

provide a complete picture of the reliability of test scores, the researcher must examine all types of reliability.

Even if a researcher examines every type of reliability, the results are incomplete and hard to interpret. First, the reliability results are incomplete because they do not consider the interaction of these factors. To what extent do ratings change over time? Do some raters score some items more harshly? Is the change in ratings over time consistent across items? Thus, classical test score theory does not take into account two-way and three-way interactions between items, time, and raters. Second, the reliability results are hard to interpret because each coefficient is given separately. If internal consistency is .91, test–retest reliability is .85, and interrater reliability is .82, then what proportion of observed score variance is a result of true differences between participants and what proportion is a result of these three sources of random error?

To address these issues, researchers can use more sophisticated mathematical models, which are based on a multifactor repeated measures analysis of variance (ANOVA). First, researchers can conduct a study to examine the influence of all these factors on test scores. Second, researchers can calculate generalizability coefficients to take into account the number of items, times, and raters that will be used when collecting data to make decisions in an applied context.

*Kimberly A. Barchard*

***See also*** Classical Test Theory; Coefficient Alpha; Exploratory Factor Analysis; Generalizability Theory; Interrater Reliability; Intraclass Correlation; KR-20; Reliability; Structural Equation Modeling; Test–Retest Reliability

**Further Readings**

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78,* 98–104.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Orlando, FL: Harcourt Brace Jovanovich.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology, 16,* 137–163.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2,* 151–160.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Raykov, R., & Shrout, P. E. (2002). Reliability of scales with general structure: Point and interval estimation with structural equation modeling approach. *Structural Equation Modeling: A Multidisciplinary Journal, 9,* 195–212.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8,* 350–353.

**Websites**

International Personality Item Pool: http://ipip.ori.org

# Internal Validity

Internal validity refers to the accuracy of statements made about the causal relationship between two variables, namely, the manipulated (treatment or independent) variable and the measured variable (dependent). Internal validity claims are not based on the labels a researcher attaches to variables or how they are described but, rather, to the procedures and operations used to conduct a research study, including the choice of design and measurement of variables. Consequently, internal validity is relevant to the topic of research methods. In the next three sections, the procedures that support causal inferences are introduced, the threats to internal validity are outlined, and methods to follow to increase the internal validity of a research investigation are described.

## Causal Relationships Between Variables

When two variables are correlated or found to covary, it is reasonable to ask the question of whether there is a direction in the relationship. Determining whether there is a causal relationship

between the variables is often done by knowing the time sequence of the variables; that is, whether one variable occurred first followed by the other variable. In randomized experiments, where participants are randomly assigned to treatment conditions or groups, knowledge of the time sequence is often straightforward because the treatment variable (independent) is manipulated before the measurement of the outcome variable (dependent). Even in quasi-experiments, where participants are *not* randomly assigned to treatment groups, the investigator can usually relate some of the change in pre-post test measures to group membership. However, in observational studies where variables are not being manipulated, the time sequence is difficult, if not impossible, to disentangle.

One might think that knowing the time sequence of variables is often sufficient for ascertaining internal validity. Unfortunately, time sequence is not the only important aspect to consider. Internal validity is also largely about ensuring that the causal relationship between two variables is direct and not mitigated by a third variable. A third, uncontrolled, variable can function to make the relationship between the two other variables appear stronger or weaker than it is in real life. For example, imagine that an investigator decides to investigate the relationship between class size (treatment variable) and academic achievement (outcome variable). The investigator recruits school classes that are considered large (with more than 20 students) and classes that are considered small (with fewer than 20 students). The investigator then collects information on students' academic achievement at the end of the year to determine whether a student's achievement depends on whether he or she is in a large or small class. Unbeknownst to the investigator, however, students who are selected to small classes are those who have had behavioral problems in the previous year. In contrast, students assigned to large classes are those who have not had behavioral problems in the previous year. In other words, class size is related negatively to behavioral problems. Consequently, students assigned to smaller classes will be more disruptive during classroom instruction and will plausibly learn less than those assigned to larger classes. In the course of data analysis, if the investigator were to discover a significant relationship between class size and academic achievement, one could argue that this relationship is not direct. The investigator's discovery is a *false positive finding*. The relationship between class size and academic achievement is not direct because the students associated with classes of different sizes are not equivalent on a key variable—attentive behavior. Thus, it might not be that larger class sizes have a positive influence on academic achievement but, rather, that larger classes have a selection of students that, without behavioral problems, can attend to classroom instruction.

The third variable can threaten the internal validity of studies by leading to false positive findings or false negative findings (i.e., not finding a relationship between variables A and B because of the presence of a third variable, C, that is diminishing the relationship between variables A and B). There are many situations that can give rise to the presence of uncontrolled third variables in research studies. In the next section, threats to internal validity are outlined. Although each threat is discussed in isolation, it is important to note that many of these threats can simultaneously undermine the internal validity of a research study and the accuracy of inferences about the causality of the variables involved.

## Threats to Internal Validity

### *1. History*

An event (e.g., a new video game), which is not the treatment variable of interest, becomes accessible to the treatment group but not the comparison group during the pre- and posttest time interval. This event influences the observed effect (i.e., the outcome, dependent variable). Consequently, the observed effect cannot be attributed exclusively to the treatment variable (thus threatening internal validity claims).

### *2. Maturation*

Participants develop or grow in meaningful ways during the course of the treatment (between the pretest and posttest). The developmental change in participants influences the observed effect, and so now the observed effect cannot be solely attributed to the treatment variable.

*3. Testing*

In the course of a research study, participants might be required to respond to a particular instrument or test multiple times. The participants become familiar with the instrument, which enhances their performance and the observed effect. Consequently, the observed effect cannot be solely attributed to the treatment variable.

*4. Instrumentation*

The instrument used as a pretest to measure participants is not the same as the instrument used for the posttest. The differences in test type could influence the observed effect; for example, the metric used for the posttest could be more sensitive to changes in participant performance than the metric used for the pretest. The change in metric and not the treatment variable of interest could influence the observed effect.

*5. Statistical Regression*

When a pretest measure lacks reliability and participants are assigned to treatment groups based on pretest scores, any gains or losses indicated by the posttest might be misleading. For example, participants who obtained low scores on a badly designed pretest are likely to perform better on a second test such as the posttest. Higher scores on the posttest might give the appearance of gains resulting from the manipulated treatment variable but, in fact, the gains are largely caused by the inaccurate measure originally provided by the pretest.

*6. Mortality*

When participants are likely to drop out more often from one treatment group in relation to another (the control), the observed effect cannot be attributed solely to the treatment variable. When groups are not equivalent, any observed effect could be caused by differences in the composition of the groups and not the treatment variable of interest.

*7. Selection*

Internal validity is compromised when one treatment group differs systematically from another group on an important variable. In the example described in the previous section, the two groups of class sizes differed systematically in the behavioral disposition of students. As such, any observed effect could not be solely attributed to the treatment variable (class size). Selection is a concern when participants are not randomly assigned to groups. This category of threat can also interact with other categories to produce, for example, a selection-history threat, in which treatment groups have distinct local events occurring to them as they participate in the study, or a selection-maturation threat, in which treatment groups have distinct maturation rates that are unrelated to the treatment variable of interest.

*8. Ambiguity About Direction of Causal Influence*

In correlation studies that are cross-sectional, meaning that variables of interest have not been manipulated and information about the variables are gathered at one point in time, establishing the causal direction of effects is unworkable. This is because the temporal precedence among variables is unclear. In experimental studies, in which a variable has been manipulated, or in correlation studies, where information is collected at multiple time points so that the temporal sequence can be established, this is less of a threat to internal validity.

*9. Diffusion of Treatment Information*

When a treatment group is informed about the manipulation and then happens to share this information with the control group, this sharing of information could nullify the observed effect. The sharing of details about the treatment experience with control participants effectively makes the control group similar to the treatment group.

*10. Compensatory Equalization of Treatments*

This is similar to the threat described in number 9. In this case, however, what nullifies the effect of the treatment variable is not the communication between participants of different groups but, rather, administrative concerns about the inequality of the treatment groups. For example, if an experimental school receives extra funds to implement an innovative curriculum, the control school might be given similar funds and encouraged to

develop a new curriculum. In other words, when the treatment is considered desirable, there might be administrative pressure to compensate the control group, thereby undermining the observed effect of the treatment.

### *11. Rivalry Between Treatment Conditions*

Similar to the threat described in number 9, threat number 10 functions to nullify differences between treatment groups and, thus, an observed effect. In this case, when participation in a treatment versus control group is made public, control participants might work extra hard to outperform the treatment group. Had participants not been made aware of their group membership, an observed effect might have been found.

### *12. Demoralization of Participants Receiving Less Desirable Treatments*

This last threat is similar to the one described in number 11. In this case, however, when treatment participation is made public and the treatment is highly desirable, control participants might feel resentful and disengage with the study's objective. This could lead to large differences in the outcome variable between the treatment and control groups. However, the observed outcome might have little to do with the treatment and more to do with participant demoralization in the control group.

## Establishing Internal Validity

Determining whether there is a causal relationship between variables, A and B, requires that the variables covary, the presence of one variable preceding the other (e.g., A $\rightarrow$ B), and ruling out the presence of a third variable, C, which might mitigate the influence of A on B. One powerful way to enhance internal validity is to randomly assign sample participants to treatment groups or conditions. By randomly assigning, the investigator can guarantee the probabilistic equivalence of the treatment groups before the treatment variable is administered. That is, any participant biases are equally distributed in the two groups. If the sample participants cannot be randomly assigned, and the investigator must work with intact groups, which is often the case in field research, steps must be taken to ensure that the groups are equivalent on key variables. For example, if the groups are equivalent, one would expect both groups to score similarly on the pretest measure. Furthermore, one would inquire about the background characteristics of the students—Are there equal distributions of boys and girls in the groups? Do they come from comparable socioeconomic backgrounds? Even if the treatment groups are comparable, efforts should be taken to not publicize the nature of the treatment one group is receiving relative to the control group so as to avoid threats to internal validity involving diffusion of treatment information, compensatory equalization of treatments, rivalry between groups, and demoralization of participants that perceive to be receiving the less desirable treatment. Internal validity checks are ultimately designed to bolster confidence in the claims made about the causal relationship between variables; as such, internal validity is concerned with the integrity of the design of a study for supporting such claims.

*Jacqueline P. Leighton*

***See also*** Cause-and-Effect; Control Variables; Quasi-Experimental Design; Random Assignment; True Experimental Design

**Further Readings**

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton-Mifflin.

Keppel, G., & Wickens, T. D. (2002). *Design and analysis: A researcher's handbook* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis* (2nd ed.). Boston: McGraw-Hill.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton-Mifflin.

# Internet-Based Research Method

Internet-based research method refers to any research method that uses the Internet to collect

data. Most commonly, the Web has been used as the means for conducting the study, but e-mail has been used as well. The use of e-mail to collect data dates back to the 1980s while the first uses of the Web to collect data started in the mid-1990s. Whereas e-mail is principally limited to survey and questionnaire methodology, the Web, with its ability to use media, has the ability to execute full experiments and implement a wide variety of research methods. The use of the Internet offers new opportunities for access to participants allowing for larger and more diverse samples. However, this new access to participants comes at the cost of a great deal of loss of control of the research environment. Although this loss of control can be problematic in correlational designs, it can be devastating in experimental designs were environmental control can be everything. What makes this method so intriguing is the fact that valid results have been obtained in many studies using this methodology.

Even though the use of e-mail is an interesting research method, it is little used at this time. It is probably because e-mail offers few if any advantages over the Web as a research environment, and it cannot perform any data collection methods that are not possible with the Web. As a result, the rest of this entry focuses on the use of the Web to collect psychological data.

## Some General Terms

To ease discussion, some general terms need to be introduced as they are unique to research on the Internet. These terms are shown in Table 1.

## Development

The use of the Web for data collection primarily required the development of *forms* in Web pages. Forms are so ubiquitous now that it is hard to realize that forms are not an original part of the Web, which was primarily a novel way of moving between and within documents via links (hypertext). The next development was the ability to embed images in the browser known as Mosaic. However, until forms were developed, there was no means to gain information from the user. With the development of forms, and the simultaneous means to store the information from the forms, it was possible to ask readers of Web pages to submit information, and then these data could be sent to the server and stored. This development occurred in the mid-1990s and it was not long before psychological researchers took advantage of this possibility to present stimuli to participants and collect their responses, thereby conducting experiments over the Internet.

From that point in time, the use of the Internet for research has grown dramatically. In the mid-1990s, only a handful of studies were posted each year; now, hundreds of new studies are posted each year as shown on websites like *Psychological Research on the Net*. The largest proportion of studies is in the area of social and cognitive psychology, although studies have been conducted in most areas of psychology including emotions, mental health, health psychology, perception, and even biopsychology.

## Examples

The first publication of actual formal psychological experiments can be found in the 1997 papers

**Table 1** General Terms to Understand Internet Research

| *Term* | *Definition* |
|---|---|
| Browser | The software used to read Web pages. Popular examples include Internet Explorer and Firefox. |
| Server | The machine/software that host the Web page. When the browser asks for a Web page, it asks for the server to deliver the page. |
| Client | The machine the person reading the Web page is on. The browser resides on the client machine. |
| Server side | Operations that take place on the server. For example, when the data are stored, the operations that store the date are server side. |
| Client side | Operations that take place on the client machine. For example, the browser that interprets a Web page file and makes it so you can read it is client side. |
| Forms | Elements of Web pages that allow for the user to input information that can be sent to the server. |

by John Krantz, Jody Ballard, and Jody Scher, as well as by Ulf Reips. One main experiment performed by Krantz, Ballard, and Scher was a within-subject experiment examining preferences for different weighted female drawings. The experiment was a 3 × 9 design with independent variables on weight and shoulder-to-hip proportions. The experiment was conducted both under traditional laboratory conditions and on the Web. The participants gave ratings of preference for each figure using a magnitude estimation procedure. The use of the within-subject methodology and a magnitude estimation procedure allowed for a detailed comparison of the results found in the laboratory with the results found on the Web. First, the results were highly correlated between laboratory and Web. In addition, a regression analysis was performed on the two data sets to determine whether the data do more than move in the same direction. This regression found that the Web values are nearly identical to the values for the same condition in the laboratory; that is, the Web data essentially can be replaced by the laboratory data and vice versa. This similarity was found despite the vast difference in the ways the experiment was delivered (e.g., in the laboratory, the participants were tested in groups, and on the Web, presumably most participants ran the study singly) and differences in age range (in the laboratory, all participants are traditional college students, whereas a much greater age range was observed in the Web version).

Krantz and Reshad Dalal performed a literature review of the Web studies conducted up to the time this entry was written. The point of the review was to determine whether the Web results could be considered, at least in a general sense, valid. They examined both e-mail and Web-based research methods and compared their results with laboratory method results. In general, they found that most Web studies tended to find weaker results than in the control of the laboratory. However, the results seemed valid and even in cases where the data differed from the laboratory or field, the differences were intriguing. One study performed an e-mail version of Stanley Milgram's lost letter technique. In Milgram's study, the letters that were mailed were sent to the originally intended destination. However, the e-mails that were sent were returned to the original sender. In both cases, the easiest path to be helpful was predominantly taken. So both studies complement each other.

## Practical Issues

Next, several practical issues that can influence the data quality obtained from Web-based studies is discussed. When a Web-based study is read, it is important to understand how the experimenter handles the following factors.

### *Recruitment*

Just placing a study on the Web is not usually sufficient to get an adequately large sample to analyze. It is typical to advertise the study. There are several methods of study advertising that are used. One way is to advertise on sites that list psychological studies. The two largest are listed at the end of this entry. These sites are also well known to people interested in participating in psychological research, which makes them a useful means for participants to find research studies. These sites also come up at the top of searches for psychological experiments and related terms in search engines. Another common method is to solicit participants from discussion groups or e-mail listservs. Because these groups tend to be formed to discuss common issues, this method allows access to subpopulations that might be of interest. With the advent of social networking on the Web, social networking sites such as Facebook have also been used to recruit participants. Finally, traditional media such as radio and television can be used to recruit participants to Internet studies. It has occurred that some network news programs have found an experiment related to a show they were running and posted a link to that study on the website associated with the show. It should be noted that the Web is not a monolithic entity. Different methods of recruitment will lead to different samples. Depending on the sample needs of the study, it is often advisable to use multiple types of recruitment methods.

### *Sample Characteristics*

One enticing feature of the Web as a research environment is the ability to obtain more diverse

samples. Samples are more diverse on the Web than the comparable sample in the laboratory. However, that is not to say that the samples are truly representative. Web use is not even distributed across all population segments. It is probably wise to consider the Web population in a mode similar to early days of the telephone which, when used for sampling without attention to the population that had telephones, led to some classic mistaken conclusions in political polls.

### *Dropout*

One of the big concerns in Web-based research is the ease with which participants can leave the study. In the laboratory, it is rare for a participant to up and leave the experiment. Participants in Web-based research regularly do not complete a study, leaving the researcher with several incomplete data sets. Incomplete data can make up to 40% of a data set in some studies. There are two main concerns regarding dropout. First, if the dropout is not random but is selective in some sense, it can limit the generalizability of the results. Second, if the conditions in experiments differ in a way that causes differential dropout across conditions, this fact can introduce a confound in the experiment. This factor must be examined in evaluating the conditions. Information about the length of the study and the careful use of incentives can reduce dropout. In addition, it is possible in experiments that use many pages to measure dropout and use it as a variable in the data analysis.

### *Technical Variance*

One of the big differences between the laboratory and the Web is the loss of control over the equipment used by the participant. In the laboratory, it is typical to have the participants all use the same computer or same type of computer, control environmental conditions, and any other factor that might influence the outcome of the study. On the Web, such control is not possible. Variations include the type of computer being used, the way the person is connected to the network, the type of browser, the size of browser window the participant prefers, the version of the browser, and even what other programs might be running in the background. All of these factors, and others, add potential sources of error to the data collected over the Web. The term *technical variance* has been applied to this source of variation in experimental conditions. Many of these variations, such as browser type and version, can be collected during the experiment, allowing some assessment of the influence of these technical variations on the data. However, although rare, it is possible for users to hide or alter these values, such as altering what browser is being used.

## Ethical Issues

There have been some major discussions of the ethics of Web-based research. In particular, the lack of contact between the experimenter and participant means that it is more difficult to ensure that the nature of the study is understood. In addition, there is no way to be sure that any participant is debriefed rendering the use of deception particularly problematic on the Web. However, on the positive side, participants do feel very free to leave the study at any time, meaning that it can be more clearly assumed that the sample is truly voluntary, free of the social constraints that keep many participants in laboratory experiments when they wish to leave.

## Future Directions

As Web research becomes more widely accepted, the main future direction will be Web-based research examining new topic areas that are not possible to be done in the laboratory. To date most Web-based studies have been replications and extensions of existing research studies. Another development will be the greater use of media in the experiments. Most studies to date have been principally surveys with maybe a few images used as stimuli. The variations of monitors and lighting have rendered the use of any images, beyond the simplest, problematic. The development of better and more controlled methods of delivering images and video will allow a wider range of studies to be explored over the Web.

*John H. Krantz*

*See also* Bias; Confounding; Ethics in the Research Process; Experimental Design; Sampling

### Further Readings

Binbaum, M. H. (2000). *Psychological experiments on the internet.* San Diego, CA: Academic Press.
Krantz, J. H., Ballard, J., & Scher, J. (1997). Comparing the results of laboratory and world-wide Web samples on the determinants of female attractiveness. *Behavioral Research Methods, Instruments, & Computers, 29,* 264–269.
Krantz, J. H., & Dalal, R. (2000). Validity of Web-based psychological research. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 35–60). San Diego, CA: Academic Press.
Musch, J., & Reips, U.-D. (2000). A brief history of Web experimenting. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 61–88). San Diego, CA: Academic Press.
Reips, U.-D. (2002). Standards for Internet-based experimenting. *Experimental Psychology, 49,* 243–256.

### Websites

Psychological Research on the Net: http://psych.hanover.edu/research/exponnet.html
Web Experiment List: http://genpsylab-wexlist.unizh.ch

## Interrater Reliability

The use of raters or observers as a method of measurement is prevalent in various disciplines and professions (e.g., psychology, education, anthropology, and marketing). For example, in psychotherapy research raters might categorize verbal (e.g., paraphrase) and/or nonverbal (e.g., a head nod) behavior in a counseling session. In education, three different raters might need to score an essay response for advanced placement tests. This type of reliability is also present in other facets of modern society. For example, medical diagnoses often require a second or even third opinion from physicians. Competitions, such as Olympic figure skating, award medals based on quantitative ratings provided by a panel of judges.

Those data recorded on a rating scale are based on the subjective judgment of the rater. Thus, the generality of a set of ratings is always of concern. Generality is important in showing that the obtained ratings are not the idiosyncratic results of one person's subjective judgment. Procedure questions include the following: How many raters are needed to be confident in the results? What is the minimum level of agreement that the raters need to achieve? Is it necessary for the raters to agree exactly or is it acceptable for them to differ from one another as long as the differences are systematic? Are the data nominal, ordinal, or interval? What resources are available to conduct the interrater reliability study (e.g., time, money, and technical expertise)?

Interrater or interobserver (these terms can be used interchangeably) reliability is used to assess the degree to which different raters or observers make consistent estimates of the same phenomenon. Another term for interrater or interobserver reliability estimate is consistency estimates. That is, it is not necessary for raters to share a common interpretation of the rating scale, as long as each judge is consistent in classifying the phenomenon according to his or her own viewpoint of the scale. Interrater reliability estimates are typically reported as correlational or analysis of variance indices. Thus, the interrater reliability index represents the degree to which ratings of different judges are proportional when expressed as deviations from their means. This is not the same as interrater agreement (also known as a consensus estimate of reliability), which represents the extent to which judges make exactly the same decisions about the rated subject. When judgments are made on a numerical scale, interrater agreement generally means that the raters assigned exactly the same score when rating the same person, behavior, or object. However, the researcher might decide to define agreement as either identical ratings or ratings that differ no more than one point or as ratings that differ no more than two points (if the interest is in judgment similarity). Thus, agreement does not have to be defined as an all-or-none phenomenon. If the researcher does decide to include a discrepancy of one or two points in the definition of agreement, the chi-square value for identical agreement should also be reported. It is possible to have high interrater reliability but low interrater agreement and vice versa. The researcher must determine which form of determining rater reliability is most important for the particular study.

Whenever rating scales are being employed, it is important to pay special attention to the interrater or interobserver reliability and interrater agreement of the rating. It is essential that both the reliability and agreement of the ratings are provided before the ratings are accepted. In reporting the interrater reliability and agreement of the ratings, the researcher must describe the way in which the index was calculated.

The remainder of this entry focuses on calculating interrater reliability and choosing an appropriate approach for determining interrater reliability.

## Calculations of Interrater Reliability

For nominal data (i.e., simple classification), at least two raters are used to generate the categorical score for many participants. For example, a contingency table is drawn up to tabulate the degree of agreement between the raters. Suppose 100 observations are rated by two raters and each rater checks one of three categories. If the two raters checked the same category in 87 of the 100 observations, the percentage of agreement would be 87%. The percentage of agreement gives a rough estimate of reliability and it is the most popular method of computing a consensus estimate of interrater reliability. The calculation is also easily done by hand. Although it is a crude measure, it does work no matter how many categories are used in each observation. An adequate level of agreement is generally considered to be 70%. However, a better estimate of reliability can be obtained by using Cohen's kappa, which ranges from 0 to 1 and represents the proportion of agreement corrected for chance.

$$K = (\rho_a - \rho_c)/(1 - \rho_c),$$

where $\rho_a$ is the proportion of times the raters agree and $\rho_c$ is the proportion of agreement we would expect by chance. This formula is recommended when the same two judges perform the ratings. For Cohen's kappa, .50 is considered acceptable. If subjects are rated by different judges but the number of judges rating each observation is held constant, then Fleiss' kappa is preferred.

Consistency estimates of interrater reliability are based on the assumption that it is not necessary for the judges to share a common interpretation of the rating scale, as long as each rater is consistent in assigning a score to the phenomenon. Consistency is most used with continuous data. Values of .70 or better are generally considered to be adequate. The three most common types of consistency estimates are (1) correlation coefficients (e.g., Pearson and Spearman), (2) Cronbach's alpha, and (3) intraclass correlation.

The Pearson product-moment correlation coefficient is the most widely used statistic for calculating the degree of consistency between independent raters. Values approaching $+1$ or $-1$ indicate that the raters are following a consistent pattern, whereas values close to zero indicate that it would be almost impossible to predict the rating of one judge given the rating of the other judge. An acceptable level of reliability using a Pearson correlation is .70. Pearson correlations can only be calculated for one pair of judges at a time and for one item at a time. The Pearson correlation assumes the underlying data are normally distributed. If the data are not normally distributed, the Spearman rank coefficient should be used. For example, if two judges rate a response to an essay item from best to worst, then a ranking and the Spearman rank coefficient should be used.

If more than two raters are used, Cronbach's alpha correlation coefficient could be used to compute interrater reliability. An acceptable level for Cronbach's alpha is .70. If the coefficient is lower than .70, this means that most of the variance in the total composite score is a result of error variance and not true score variance.

The best measure of interrater reliability available for ordinal and interval data is the intraclass correlation ($R$). It is the most conservative measure of interrater reliability. $R$ can be interpreted as the proportion of the total variance in the ratings caused by variance in the persons or phenomena being rated. Values approaching the upper limit of $R(1.00)$ indicate a high degree of reliability, whereas an $R$ of 0 indicates a complete lack of reliability. Although negative values of $R$ are possible, they are rarely observed; when they are observed, they imply judge $\times$ item interactions. The more $R$ departs from 1.00, the less reliable are the judge's ratings. The minimal acceptable level of $R$ is considered to be .60. There is more than one formula available for intraclass correlation. To select the appropriate formula, the investigator must decide

(a) whether the mean differences in the ratings of the judges should be considered rater error and (b) whether he or she is more concerned with the reliability of the average rating of all the judges or the average reliability of the individual judge.

The two most popular types of measurement estimates of interrater reliability are (a) factor analysis and (b) the many-facets Rasch model. The primary assumption for the measurement estimates is that all the information available from all the judges (including discrepant ratings) should be used when calculating a summary score for each respondent. Factor analysis is used to determine the amount of shared variance in the ratings. The minimal acceptable level is generally 70% of the explained variance. Once the interrater reliability is established, each subject will receive a summary score based on his or her loading on the first principal component underlying the ratings. Using the many-facets Rasch model, the ratings between judges can be empirically determined. Also the difficulty of each item, as well as the severity of all judges who rated each item, can be directly compared. In addition, the facets approach can determine to what degree each judge is internally consistent in his or her ratings (i.e., an estimate of intrarater reliability). For the many-facets Rasch model, the acceptable rater values are greater than .70 and less than 1.3.

## Choosing an Approach

There is no "best" approach for calculating interrater or interobserver reliability. Each approach has its own assumptions and implications as well as its own strengths and weaknesses. The percentage of agreement approach is affected by chance. Low prevalence of the condition of interest will affect kappa and correlations will be affected by low variability (i.e., attenuation) and distribution shape (normality or skewed). Agreement estimates of interrater reliability (percent agreement, Cohen's kappa, Fleiss' kappa) are generally easy to compute and will indicate rater disparities. However, training raters to come to an exact consensus will require considerable time and might or might not be necessary for the particular study.

Consistency estimates of interrater reliability (e.g., Pearson product-moment and Spearman rank correlations, Cronbach's alpha coefficient, and intraclass correlation coefficients) are also fairly simple to compute. The greatest disadvantage to using these statistical techniques is that they are sensitive to the distribution of the data. The more the data depart from a normal distribution, the more attenuated the results.

The measurement estimates of interrater reliability (e.g., factor analysis and many-facets Rasch measurement) can work with multiple judges, can adjust summary scores for rater severity, and can allow for efficient designs (e.g., not all raters have to judge each item or object). However, the measurement estimates of interrater reliability require expertise and considerable calculation time.

Therefore, as noted previously, the best technique will depend on the goals of the study, the nature of the data (e.g., degree of normality), and the resources available. The investigator might also improve reliability estimates with additional training of raters.

*Karen D. Multon*

***See also*** Cohen's Kappa; Correlation; Instrumentation; Intraclass Correlation; Pearson Product-Moment Correlation Coefficient; Reliability; Spearman Rank Order Correlation

### Further Readings

Bock, R., Brennan, R. L., & Muraki, E. (2002). The information in multiple ratings. *Applied Psychological Measurement, 26,* 364–375.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37–46.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76,* 378–382.

Linacre, J. M. (1994). *Many-facet Rasch measurement.* Chicago: MESA Press.

Snow, A. L., Cook, K. F., Lin, P. S., Morgan, R. O., & Magaziner, J. (2005). Proxies and other external raters: Methodological considerations. *Health Services Research, 40,* 1676–1693.

Stemler, S. E., & Tsai, J. (2008). Best practices in interrater reliability: Three common approaches. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 29–49). Thousand Oaks, CA: Sage.

Tinsley, H. E. A., & Weiss, D. J. (1975). Interrater reliability and agreement of subjective judgements. *Journal of Counseling Psychology, 22,* 358–376.

# Interval Scale

Interval scale refers to the level of measurement in which the attributes composing variables are measured on specific numerical scores or values and there are equal distances between attributes. The distance between any two adjacent attributes is called an *interval,* and intervals are always equal.

There are four scales of measurement, which include nominal, ordinal, interval, and ratio scales. The ordinal scale has logically rank-ordered attributes, but the distances between ranked attributes are not equal or are even unknown. The equal distances between attributes on an interval scale differ from an ordinal scale. However, interval scales do not have a "true zero" point, so statements about the ratio of attributes in an interval scale cannot be made. Examples of interval scales include temperature scales, standardized tests, the Likert scale, and the semantic differential scale.

## Temperature Scales and Standardized Tests

Temperature scales including the Fahrenheit and Celsius temperature scales are examples of an interval scale. For example, the Fahrenheit temperature scale in which the difference between 25° and 30° is the same as the difference between 80° and 85°. In the Celsius temperature scales, the distance between 16° and 18° is the same as that between 78° and 80°.

However, 60°F is not twice as hot as 30°F. Similarly, –40°C is not twice as cold as –20°C. This is because both Fahrenheit and Celsius temperature scales do not have a "true zero" point. The zero points in the Fahrenheit and Celsius temperature scales are arbitrary—in both scales, 0° does not mean the lack of heat nor cold.

In contrast, the Kelvin temperature scale is based on a "true zero" point. The zero point of the Kelvin temperature scale, which is equivalent to –459.67°F or –273.15°C is considered the lowest possible temperature of anything in the universe. In the Kelvin temperature scale, 400 K is twice as hot as 200 K, and 100 K is twice as cold as 200 K. The Kelvin temperature scale is not an example of interval scale but that of ratio scale.

Standardized tests, including Intelligence Quotient (IQ), Scholastic Assessment Test (SAT), Graduate Record Examination (GRE), Graduate Management Admission Test (GMAT), and Miller Analogies Test (MAT) are also examples of an interval scale. For example, in the IQ scale, the difference between 150 and 160 is the same as that between 80 and 90. Similarly, the distance in the GRE scores between 350 and 400 is the same as the distance between 500 and 550.

Standardized tests are not based on a "true zero" point that represents the lack of intelligence. These standardized tests do not even have a zero point. The lowest possible score for these standardized tests is not zero. Because of the lack of a "true zero" point, standardized tests cannot make statements about the ratio of their scores. Those who have an IQ score of 150 are not twice as intelligent as those who have an IQ score of 75. Similarly, such a ratio cannot apply to other standardized tests including SAT, GRE, GMAT, or MAT.

## Likert Scales

One example of interval scale measurement that is widely used in social science is the Likert scale. In experimental research, particularly in social sciences, there are measurements to capture attitudes, perceptions, positions, feelings, thoughts, or points of view of research participants. Research participants are given questions and they are expected to express their responses by choosing one of five or seven rank-ordered response choices that is closest to their attitudes, perceptions, positions, feelings, thoughts, or points of view.

An example of the Likert scales that uses a 5-point scale is as follows:

> How satisfied are you with the neighborhood where you live?
>
> - Very satisfied
> - Somewhat satisfied
> - Neither satisfied nor dissatisfied
> - Somewhat dissatisfied
> - Very dissatisfied

Some researchers argue that such responses are not interval scales because the distance between

attributes are not equal. For example, the difference between *very satisfied* and *somewhat satisfied* might not be the same as that between *neither satisfied nor dissatisfied* and *somewhat dissatisfied*.

Each attribute in the Likert scales is given a number. For the previous example, *very satisfied* is 5, *somewhat satisfied* is 4, *neither satisfied nor dissatisfied* is 3, *somewhat dissatisfied* is 2, and *very dissatisfied* is 1. The greater number represents the higher degree of satisfaction of respondents of their neighborhood. Because of such numbering, there is now equal distance between attributes. For example, the difference between *very satisfied* (5) and *somewhat satisfie*d (4) is the same as the difference between *neither satisfied nor dissatisfied* (3) and *somewhat dissatisfied* (2). However, the Likert scale does not have a "true zero" point, as shown in the previous example, so that statements about the ratio of attributes in the Likert scale cannot be made.

### Semantic Differential Scale

Another interval scale measurement is the semantic differential scale. Research respondents are given questions and also semantic differential scales, usually 7-point or 5-point response scales, as their response choices. Research respondents are expected to choose 1 scale out of 7 or 5 semantic differential scales that is closest to their condition or perception.

An example of the semantic differential scales that uses a 7-point scale is as follows:

> How would you rate the quality of the neighborhood where you live?

| Excellent | | | | | | Poor |
|---|---|---|---|---|---|---|
| 7 | 6 | 5 | 4 | 3 | 2 | 1 |

Research respondents who rate the quality of their neighborhood as excellent should choose "7" and those who rate the quality of their neighborhood as poor should choose "1." Similar to the Likert scale, there is equal distance between attributes in the semantic differential scales, but there is no "true zero" point.

*Deden Rukmana*

***See also*** Ordinal Scale; Ratio Scale

**Further Readings**

Babbie, E. (2007). *The practice of social research* (11th ed.). Belmont, CA: Thomson and Wadsworth.

Dillman, D. A. (2007). *Mail and internet surveys: The tailored design method* (2nd ed.). New York: Wiley.

Keyton, J. (2006). *Communication research: Asking questions, finding answers* (2nd ed.). Boston: McGraw-Hill.

University Corporation for Atmospheric Research. (2001). W*indows to the universe: Kelvin scale*. Retrieved December 14, 2008, from http://www.windows.ucar.edu/cgi-bin/tour_def/earth/Atmosphere/temperature/kelvin.html

## INTERVENTION

Intervention research examines the effects of an intervention on an outcome of interest. The primary purpose of intervention research is to engender a desirable outcome for individuals in need (e.g., reduce depressive symptoms or strengthen reading skills). As such, intervention research might be thought of as differing from prevention research, where the goal is to prevent a negative outcome from occurring, or even from classic laboratory experimentation, where the goal is often to support specific tenets of theoretical paradigms. Assessment of an intervention's effects, the sine qua non of intervention research, varies according to study design, but typically involves both statistical and logical inferences.

The hypothetical intervention study presented next is used to illustrate important features of intervention research. Assume a researcher wants to examine the effects of parent training (i.e., intervention) on disruptive behaviors (i.e., outcome) among preschool-aged children. Of 40 families seeking treatment at a university-based clinic, 20 families were randomly assigned to an intervention condition (i.e., parent training) and the remaining families were assigned to a (wait-list) control condition. Assume the intervention was composed of six, 2-hour weekly therapy sessions with the parent(s) to strengthen theoretically identified parenting practices (e.g., effective discipline strategies) believed to reduce child disruptive behaviors. Whereas parents assigned to the intervention condition attended sessions, parents assigned to the

control condition received no formal intervention. In the most basic form of this intervention design, data from individuals in both groups are collected at a single baseline (i.e., preintervention) assessment and at one follow-up (i.e., postintervention) assessment.

## Assessing the Intervention's Effect

In the parenting practices example, the first step in assessing the intervention's effect involves testing for a statistical association between intervention group membership (intervention vs. control) and the identified outcome (e.g., reduction in temper tantrum frequency). This is accomplished by using an appropriate inferential statistical procedure (e.g., an independent-samples $t$ test) coupled with an effect size estimate (e.g., Cohen's $d$), to provide pertinent information regarding both the statistical significance and strength (i.e., the amount of benefit) of the intervention–outcome association.

Having established an intervention–outcome association, researchers typically wish to ascertain whether this association is causal in nature (i.e., that the intervention, not some other factor, caused the observed group difference). This more formidable endeavor of establishing an "intervention to outcome" causal connection is known to social science researchers as establishing a study's internal validity—the most venerable domain of the renowned Campbellian validity typology. Intervention studies considered to have high internal validity have no (identified) plausible alternative explanations (i.e., internal validity threats) for the intervention–outcome association. As such, the most parsimonious explanation for the results is that the intervention caused the outcome.

## Random Assignment in Intervention Research

The reason random assignment is a much-heralded design feature is its role in reducing the number of alternative explanations for the intervention–outcome association. In randomized experiments involving a no-treatment control, the control condition provides incredibly important information regarding what would have happened to the intervention participants had they not been exposed to the intervention. Because random assignment precludes systematic pretest group differences (as the groups are probabilistically equated on all measured and unmeasured characteristics), it is unlikely that some other factor resulted in postintervention group differences. It is worth noting that this protection conveyed by random assignment can be undone once the study commences (e.g., by differential attrition or participant loss). It is also worth noting that quasi-experiments or intervention studies that lack random assignment to condition are more vulnerable to internal validity threats. Thoughtful design and analysis of quasi-experiments typically involve identifying several plausible internal validity threats a priori and incorporating a mixture of design and statistical controls that attempt to rule out (or render implausible) the influence of these threats.

## Other Things to Consider

Thus far, this discussion has focused nearly exclusively on determining whether the intervention worked. In addition, intervention researchers often examine whether certain subgroups of participants benefited more from exposure to the intervention than did other subgroups. In the parenting example, one might find that parents with a single child respond more favorably to the intervention than do parents with multiple children. Identifying this subgroup difference might aid researchers in modifying the intervention to make it more effective for parents with multiple children. This additional variable (in this case, the subgroup variable) is referred to as an intervention moderator. The effects of intervention moderators can be examined by testing statistical interactions between intervention group membership and the identified moderator.

Intervention researchers should also examine the processes through which the intervention produced changes in the outcome. Examining these process issues typically requires the researcher to construct a conceptual roadmap of the intervention's effects. In other words, the researcher must specify the paths followed by the intervention in affecting the outcomes. These putative paths are referred to as intervention mediators. In the parenting example, these paths might be (a) better understanding of child behavior, (b) using more effective discipline practices, or (c) increased levels of parenting self-efficacy. Through statistical mediation analysis, researchers can test empirically

whether the intervention affected the outcome in part by its impact on the identified mediators.

*Christian DeLucia and Steven C. Pitts*

*See also* External Validity; Internal Validity; Quasi-Experimental Designs; Threats to Validity; Treatment(s)

**Further Readings**

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions.* Newbury Park, CA: Sage.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin, 54,* 297–312.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Chicago: Rand McNally.

Cronbach, L. J. (1982). *Designing evaluations of educational and social programs.* San Francisco: Jossey-Bass.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis.* New York: Lawrence Erlbaum.

Reynolds, K., & West, S. G. (1988). A multiplist strategy for strengthening nonequivalent control group designs. *Evaluation Review, 11,* 691–714.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin.

# INTERVIEWING

Interviewing is an important aspect of many types of research. It involves conducting an interview—a purposeful conversation—between two people (the interviewer and the interviewee) to collect data on some particular issue. The person asking the questions is the interviewer, whereas the person providing the answers is the interviewee (i.e., respondent). Interviewing is used in both quantitative and qualitative research and spans a wide continuum of forms, moving from totally structured to totally unstructured. It can use a range of techniques including face-to-face (in-person), telephone, videophone, and e-mail. Interviewing involves several steps, namely, determining the interviewees, preparing for the interview, and conducting the interview.

## Important Issues to Consider When Conducting an Interview

### *Interviewer Characteristics and Demeanor*

Physical attributes such as age, race, gender, and voice, as well as attitudinal attributes such as friendliness, professionalism, optimism, persuasiveness, and confidence, are important attributes that should be borne in mind when selecting interviewers. Even when questions are well written, the success of face-to-face and telephone surveys are still very much dependent on the interviewer. Interviews are conducted to obtain information. However, information can only be obtained if respondents feel sufficiently comfortable in an interviewer's presence. Good interviewers have excellent social skills, show a genuine interest in getting to know their respondents, and recognize that they need to be flexible in accommodating respondents' schedules.

Research shows that interviewer characteristics can definitely affect both item response and response quality and might even affect a respondent's decision to participate in an interview. It might, therefore, be desirable in many cases to match interviewers and interviewees in an effort to solicit respondents' cooperation, especially for interviews that deal with sensitive topics (e.g., racial discrimination, inequality, health behavior, or domestic abuse) or threatening topics (e.g., illegal activities). For example, women interviewers should be used to interview domestically abused women. Matching might also be desirable in some cultures (e.g., older males to interview older males), or for certain types of groups (e.g., minority interviewers for minority groups). Additionally, matching might help to combat normative responses (i.e., responding in a socially desirable way) and might encourage respondents to speak in a more candid manner. Matching can be done on several characteristics, namely, race, age, ethnicity, and sex.

### *Interviewer Training*

When a research study is large and involves the use of many interviewers, it will require proper training, administration, coordination, and control. The purpose of interviewer training is to ensure that interviewers have the requisite skills

that are essential for the collection of high-quality, reliable, and valid data. The length of training will be highly dependent on the mode of survey execution, as well as the interviewers' experience. The International Standards Association, for example, recommends a minimum of 6 hours of training for new telephone interviewers involved in market, opinion, and social research.

Prior to entering the field, an interviewer training session should be conducted with all involved interviewers. At this session, interviewers should be given a crash course on basic research issues (e.g., importance of random sampling, reliability and validity, and interviewer-related effects). They should also be briefed on the study objectives and the general guidelines/procedures/protocol that should be followed for data collection. If a structured questionnaire is being used to collect data, it is important that the group go through the entire questionnaire, question by question, to ensure that every interviewer clearly understands the questionnaire. This should be followed by one or more demonstrations to illustrate the complete interview process. Complications and difficulties encountered during the demonstrations, along with recommendations for coping with the problems, should be discussed subsequent to the demonstration. Detailed discussion should take place on how to use probes effectively and how to quickly change "tone" if required. A pilot study should be conducted after training to identify any additional problems or issues.

### *Preparing for the Interview*

Prior to a face-to-face interview, the interviewer should either telephone or send an official letter to the interviewee to confirm the scheduled time, date, and place for the interview. One of the most popular venues for face-to-face interviews is respondents' homes; however, other venues can also be used (e.g., coffee shops, parking lots, or grocery stores). When sensitive topics are being discussed, a more private venue is desirable so that the respondent can talk candidly. Interviewers should also ensure that they are thoroughly acquainted with the questionnaire and guidelines for the interview. This will help to ensure that the interview progresses smoothly and does not deviate substantially from the estimated time required for the interview.

### *The Interview*

At the beginning of the interview, the interviewer should greet the respondent in a friendly manner, identify and introduce himself or herself, and thank the respondent for taking time to facilitate the interview. If a face-to-face interview is being conducted, the interviewer should also present the interviewee with an official letter from the institution sponsoring the research, which outlines the legitimacy of the research and other salient issues such as the interviewer's credentials. In telephone and face-to-face interviews where contact was not established in advance (e.g., a national survey), the interviewer has to try to elicit the cooperation of the potential respondent and request permission to conduct the interview. In unscheduled face-to-face interviews, many potential respondents might refuse to permit an interview for one or more of the following reasons: busy, simply not interested, language barrier, and safety concerns. In telephone interviews, the respondent might simply hang up the telephone with or without giving an excuse.

After introductions, the interviewer should then brief the respondent on the purpose of the study, explain how the study sample was selected, explain what will be done with the data, explain how the data will be reported (i.e., aggregated statistics—no personal information), and, finally, assure the respondent of anonymity and confidentiality. The interviewer should also give the respondent an idea of the estimated time required for the interview and should apprise the interviewee of his or her rights during the interview process (e.g., right to refuse to answer a question if respondent is uncomfortable with the question). If payment of any kind is to be offered, this should also be explained to the respondent.

An interviewer should try to establish good rapport with the interviewee to gain the interviewee's confidence and trust. This is particularly important in a qualitative interview. Establishing good rapport is, however, highly dependent on the interviewer's demeanor and social skills. Throughout the interview, the interviewer should try to make the conversational exchange a comfortable and

pleasant experience for the interviewee. Pleasantries and icebreakers can set the tone for the interview. At the same time, interviewers should be detached and neutral, and should refrain from offering any personal opinions. During the interview, interviewers should use a level of vocabulary that is easily understood by the respondent and should be careful about using certain gestures (this concern is applicable only to face-to-face interviews) and words because they might be considered offensive in some cultures and ethnic groups. In addition, interviewers should maintain a relaxed stance (body language communicates information) and a pleasant and friendly disposition; however, these are applicable only to face-to-face interviews. At all times, interviewers should listen attentively to the respondent and should communicate this to the respondent via paraphrases, probes, nods, and well-placed "uh-huhs" or "umms." An interviewer should not interrupt a respondent's silence that might occur because of thoughtful reflection or during an embarrassing conversation. Rather, he or she should give the respondent sufficient time to resume the conversation on his or her own, or important data might be lost. Additionally, when interviewers are dealing with sensitive issues, they should show some empathy with the respondent. When face-to-face interviews are being conducted, they should be done without an audience, if possible, to avoid distractions.

To conduct the interview, the interviewer will have to adopt a certain interview style (e.g., unstructured, semistructured, or structured). This is determined by the research goal and is explained during the training session. The style adopted has implications for the amount of control that the interviewer can exercise over people's responses. In qualitative research, interviews rely on what is referred to as an *interview guide*. An interview guide is a relatively unstructured list of general topics to be covered. Such guides permit great flexibility. In contrast, in quantitative research, an *interview schedule* is used. An interview schedule is a structured list of questions with explicit instructions. Interview schedules are standardized.

It is critically important that the interviewer follow the question wording for each question exactly to ensure consistency across interviews and to minimize the possibility of interviewer bias. Additionally, it is important that open-ended questions be recorded verbatim to minimize errors that could result from inaccurate summation. Verbatim responses will also permit more accurate coding. Throughout the interview, the interviewer should try to ensure that note taking is as unobtrusive as possible. Audio recordings should be used to back up handwritten notes if the respondent has no objection. However, it might be necessary at times to switch off the machine if the respondent seems reluctant to discuss a sensitive topic. Audio recordings offer several advantages, namely, they can verify the accuracy of handwritten notes and can be used to help interviewers to improve their interviewing techniques.

If respondents give incomplete or unambiguous answers, the interviewer should use tactful probes to elicit a more complete answer (e.g., "Anything else?" "In what ways?" "How?" "Can you elaborate a little more?"). Probes must never be used to coerce or lead a respondent; rather, they should be neutral, unbiased, and nondirective. Probes are more common with open-ended questions. However, they can also be used with closed-ended questions. For example, in a closed-ended question with a Likert scale, a respondent might give a response that cannot be classified on the scale. The interviewer could then ask: "Do you strongly agree or strongly disagree?" There are several types of probes that can be used, namely, the silent probe (remaining silent until the respondent continues), the echo probe (repeating the last sentence and asking the respondent to continue), the "uh-huh" probe (encouraging the respondent to continue), the tell-me-more probe (asking a question to get better insight), and the long question probe (making your question longer to get more detailed information).

At the conclusion of the interview, the interviewer should summarize the important points to the respondent, allow the respondent sufficient time to refine or clarify any points, reassure the respondent that the information will remain confidential, and thank the respondent for his or her time. Closure should be conducted in a courteous manner that does not convey abruptness to the interviewee. The respondent should be given the interviewer's contact information. An official follow-up thank-you letter should also be sent within 2 weeks. Immediately after the interview or as soon as possible thereafter, the interviewer

should update his or her recorded notes. This is particularly important when some form of shorthand notation is used to record notes.

### *Interview Debriefing*

Interview debriefing is important for obtaining feedback on the interview process. Debriefing can be held either in person or via telephone. The debriefing process generally involves asking all interviewers to fill out a questionnaire composed of both open-ended and closed-ended questions. A group meeting is subsequently held to discuss the group experiences. The debriefing session provides valuable insight on problematic issues that require correction before the next survey administration.

### *Interviewer Monitoring and Supervision*

To ensure quality control, interviewers should be supervised and monitored throughout the study. Effective monitoring helps to ensure that unforeseen problems are handled promptly, acts as a deterrent to interview falsification, and assists with reducing interviewer-related measurement error. Good monitoring focuses on four main areas: operational execution, interview quality, interviewer falsification, and survey design. In general, different types of monitoring are required for different interview techniques. For example, with face-to-face interviews, interviewers might be required to report to the principal investigator after the execution of every 25 interviews, to turn in their data and discuss any special problems encountered. In the case of telephone interviews, the monitoring process is generally simplified because interviews are recorded electronically, and supervisors also have an opportunity to listen to the actual interviews as they are being conducted. This permits quick feedback to the entire group on specific problems associated with issues such as (a) voice quality (e.g., enunciation, pace, and volume) and (b) adherence to interview protocol (e.g., reading verbatim scripts, using probes effectively, and maintaining neutrality).

## Types of Interviewing Techniques

Prior to the 1960s, paper-and-pencil (i.e., face-to-face) interviewing was the predominant type of interviewing technique. However, by the 1960s, telephone interviewing started to gain popularity. This was followed by computer-assisted telephone interviewing (CATI) in the 1970s, and computer-assisted personal interviewing (CAPI) and computer-assisted self-interviewing (CASI) in the 1980s. In CATI, an automated computer randomly dials a telephone number. All prompts for introduction and the interview questions are displayed on a computer screen. Once the respondent agrees to participate, the interviewer records the answers directly onto the computer. CAPI and CASI are quite similar to CATI but are used in face-to-face interviews. However, although CAPI is performed by the interviewer, with CASI, respondents either can be allowed to type all the survey responses onto the computer or can type the responses to sensitive questions and allow the interviewer to complete all other questions. Computer-assisted interviewing offers several advantages, including faster recording and elimination of bulky storage; however, these systems can be quite expensive to set up and data can be lost if the system crashes and the data were not backed up. Other modern-day interviewing techniques include videophone interviews, which closely resemble a face-to-face interview, except that the interviewer is remotely located, and e-mail interviews, which allow respondents to complete the interview at their convenience.

## Advantages and Disadvantages of Face-to-Face and Telephone Interviews

The administration of a questionnaire by an interviewer has several advantages compared with administration by a respondent. First of all, interviewer-administered surveys have a much higher response rate than self-administered surveys. The response rate for face-to-face interviews is approximately 80% to 85%, whereas for telephone interviews, it is approximately 60%. This might be largely attributable to the normal dynamics of human behavior. Many people generally feel embarrassed in being discourteous to an interviewer who is standing on their doorstep or is on the phone; however, they generally do not feel guilty about throwing out a mail survey as soon as it is received. Second, interviewing might help to

reduce "do not know" responses because the interviewer can probe to get a more specific answer. Third, an interviewer can clarify confusing questions. Finally, when face-to-face interviews are conducted, the interviewer can obtain other useful information, such as the quality of the dwelling (if conducted in the respondent's home), respondent's race, and respondent reactions.

Notwithstanding, interviewer-administrated surveys also have several disadvantages, namely, (a) respondents have to give real-time answers, which means that their responses might not be as accurate; (b) interviewers must have good social skills to gain respondents' cooperation and trust; (c) improper administration and interviewer characteristics can lead to interviewer-related effects, which can result in measurement error, and (d) the cost of administration is considerably higher (particularly for face-to-face interviews) compared with self-administered surveys.

### Cost Considerations

The different interviewing techniques that can be used in research all have different cost implications. Face-to-face interviews are undoubtedly the most expensive of all techniques because this procedure requires more interviewers (ratio of face-to-face to telephone is approximately 4:1), more interview time per interview (approximately 1 hour), more detailed training of interviewers, and greater supervisor and coordination. Transportation costs are also incurred with this technique. Telephone interviews are considerably cheaper—generally about half the cost. With this procedure, coordination and supervision are much easier—interviewers are generally all located in one room, printing costs are reduced, and sampling selection cost is less because samples can be selected using random-digit dialing. These cost reductions greatly outweigh the cost associated with telephone calls. Despite the significantly higher costs of face-to-face interviews, this method might still be preferred for some types of research because response rates are generally higher and the quality of the information obtained might be of a substantially higher quality compared with a telephone interview, which is quite impersonal.

### Interviewer-Related Errors

The manner in which interviews are administered, as well as an interviewer's characteristics, can often affect respondents' answers, which can lead to measurement error. Such errors are problematic, particularly if they are systematic, that is, when an interviewer makes similar mistakes across many interviews. Interviewer-related errors can be decreased through carefully worded questions, interviewer–respondent matching, proper training, continuous supervision or monitoring, and prompt ongoing feedback.

*Nadini Persaud*

***See also*** Debriefing; Ethnography; Planning Research; Protocol; Qualitative Research; Survey; Systematic Error

#### Further Readings

Babbie, E. (2005). *The basics of social research* (3rd ed.). Belmont, CA: Thomson/Wadsworth.

Erlandson, D. A., Harris, E. L., Skipper, B. L., & Allen, S. D. (1993). *Doing naturalistic inquiry: A guide to methods*. Newbury Park, CA: Sage.

Ruane, J. M. (2005). *Essentials of research methods: A guide to social sciences research*. Oxford, UK: Blackwell Publishing.

Schutt, R. K. (2001). *Investigating the social world: The process and practice of research* (3rd ed.). Thousand Oaks, CA: Pine Forge.

Weisberg, H. F., Krosnick, J. A., & Bowen, B. D. (1996). *An introduction to survey, research, polling and data analysis* (3rd ed.). Thousand Oaks, CA: Sage.

## Intraclass Correlation

The words *intraclass correlation* (ICC) refer to a set of coefficients representing the relationship between variables of the same class. Variables of the same class share a common metric and variance, which generally means that they measure the same thing. Examples include twin studies and two or more raters evaluating the same targets. ICCs are used frequently to assess the reliability of raters. The Pearson correlation coefficient usually relates measures of different classes, such as height

**Table 1** Data on Sociability of Dyads of Gay Couples

| *Couple* | *Partner 1* | *Partner 2* | *Couple* | *Partner 1* | *Partner 2* | *Couple* | *Partner 1* | *Partner 2* |
|---|---|---|---|---|---|---|---|---|
| 1 | 18 | 15 | 6 | 30 | 21 | 11 | 40 | 43 |
| 2 | 21 | 22 | 7 | 32 | 28 | 12 | 41 | 28 |
| 3 | 24 | 30 | 7 | 35 | 37 | 13 | 42 | 37 |
| 4 | 25 | 24 | 9 | 35 | 31 | 14 | 44 | 48 |
| 5 | 26 | 29 | 10 | 38 | 25 | 15 | 48 | 50 |

and weight or stress and depression, and is an interclass correlation.

Most articles on ICC focus on the computation of different ICCs and their tests and confidence limits. This entry focuses more on the uses of several different ICCs.

The different ICCs can be distinguished along several dimensions:

- One-way or two-way designs
- Consistency of order of rankings by different judges, or agreement on the levels of the behavior being rated
- Judges as a fixed variable or as a random variable
- The reliability of individual ratings versus the reliability of mean ratings over several judges

## One-Way Model

Although most ICCs involve two or more judges rating $n$ objects, the one-way models are different. A theorist hypothesizing that twins or gay partners share roughly the same level of sociability would obtain sociability data on both members of 15 gay couples from a basic sociability index. A Pearson correlation coefficient is not appropriate for these data because the data are exchangeable within couples—there is no logical reason to identify one person as the first member of the couple and the other as the second. The design is best viewed as a one-way analysis of variance with "couple" as the independent variable and the two measurements within each couple as the observations. Possible data are presented in Table 1. With respect to the dimensions outlined previously, this is a one-way design. Partners within a couple are exchangeable, and thus a partner effect would have no meaning. Because there is no partners effect, an ICC for consistency cannot be obtained, but only an ICC can be obtained for agreement. Within a couple, partner is a fixed variable—someone's partner can not be randomly select. Finally, there is no question about averaging across partners, so the reliability of an average is not relevant. (In fact, "reliability" is not really the intent.)

Table 2 gives the expected mean squares for a one-way analysis of variance. The partner effect can not be estimated separately from random error.

If each member of a couple had nearly the same score, there would be little within-couple variance, and most of the variance in the experiment would be a result of differences between couples. If members of a dyad differed considerably, the within-couple variance would be large and predominate. A measure of the degree of relationship represents the proportion ($\rho$) of the variance that is between couple variance. Therefore,

$$\rho_{ICC} = \frac{\sigma_C^2}{\sigma_C^2 + \sigma_e^2}.$$

The appropriate estimate for $\rho_{ICC}$, using the obtained mean squares ($MS$), would be

$$r_{ICC} = \frac{MS_{\text{couple}} - MS_{\text{w/in}}}{MS_{\text{couple}} + (k-1)MS_{\text{w/in}}}.$$

For this sample data, the analysis of variance summary table is shown in Table 3.

**Table 2** Expected Mean Squares for One-Way Design

| *Source* | *df* | *E(MS)* |
|---|---|---|
| Between couple | $n - 1$ | $k\sigma_C^2 + \sigma_e^2$ |
| Within couple | $n(k - 1)$ | $\sigma_e^2$ |
| Partner error | $k$ | — |
| | $(n - 1)(k - 1)$ | — |

**Table 3** Summary Table for One-Way Design

| *Source* | *df* | *Sum Sq* | *Mean Sq* | *F value* |
|---|---|---|---|---|
| Between couple | 14 | 2304.87 | 164.63 | 8.74 |
| Within couple | 15 | 282.50 | 18.83 | |
| Total | 29 | 2587.37 | | |

$$\mathrm{ICC}_1 = \frac{MS_{\text{couple}} - MS_{\text{w/in}}}{MS_{\text{couple}} + (k-1)MS_{\text{w/in}}}$$

$$= \frac{164.63 - 18.83}{164.63 + 18.83} = \frac{145.80}{183.46} = .795.$$

A test of the null hypothesis that $\rho_{\text{ICC}} = 0$ can be taken directly from the $F$ for couples, which is 8.74 on $(n-1) = 14$ and $n(k-1) = 15$ degrees of freedom ($df$).

This $F$ can then be used to create confidence limits on $\rho_{ICC}$ by defining

$$F_L = F_{\text{obs}}/F_{.975} = 8.74/2.891 = 3.023$$

$$F_U = F_{\text{obs}} \times F_{.975} = 8.74 \times 2.949 = 25.774.$$

For $F_L$, critical value is taken at $\alpha = .975$ for $(n - 1)$ and $n(k - 1)$ degrees of freedom, but for $F_U$, the degrees of freedom are reversed to obtain the critical value at $\alpha = .975$ for $n(k - 1)$ and $(n - 1)$.

The confidence interval is now given by

$$\frac{F_L - 1}{F_L + (k-1)} \le \rho \frac{F_U - 1}{F_U + (k-1)}$$

$$\frac{3.023 - 1}{3.023 + 1} \le \rho \le \frac{25.774 - 1}{25.774 + 1}.$$

$$.503 \le \rho \le .925$$

Not only are members of the same couple similar in sociability, but the ICC is large given the nature of the dependent variable.

## Two-Way Models

The previous example pertains primarily to the situation with two (or more) exchangeable measurements of each class. Two-way models usually involve different raters rating the same targets, and it might make sense to take rater variance into account.

A generic set of data can be used to illustrate the different forms of ICC. Suppose that four raters rate the compatibility of 15 married couples based on observations of a session in which couples are asked to come to a decision over a question of importance to both of them. Sample data are shown in Table 4.

**Table 4** Ratings of Compatibility for 15 Couples by Four Raters

| | *Raters* | | | |
|---|---|---|---|---|
| *Couples* | *A* | *B* | *C* | *D* |
| 1 | 15 | 18 | 15 | 18 |
| 2 | 22 | 25 | 20 | 26 |
| 3 | 18 | 15 | 10 | 23 |
| 4 | 10 | 7 | 12 | 18 |
| 5 | 25 | 22 | 20 | 30 |
| 6 | 23 | 28 | 21 | 30 |
| 7 | 30 | 25 | 20 | 27 |
| 8 | 19 | 21 | 14 | 26 |
| 9 | 10 | 12 | 14 | 16 |
| 10 | 16 | 19 | 15 | 12 |
| 11 | 14 | 18 | 11 | 19 |
| 12 | 23 | 28 | 25 | 22 |
| 13 | 29 | 21 | 23 | 32 |
| 14 | 18 | 18 | 12 | 17 |
| 15 | 17 | 23 | 14 | 15 |

### *Factors to Consider Before Computation*

#### Mixed Versus Random Models

As indicated earlier, there are several decisions to make before computing an ICC. The first is whether raters are a fixed or a random variable. Raters would be a fixed variable if they are the graduate assistants who are being trained to rate couple compatibility in a subsequent experiment. These are the only raters of concern. (The model would be a mixed model because we always assume that the targets of the ratings are sampled randomly.) Raters would be a random variable if they have been drawn at random to assess whether a rating scale we have developed can be used reliably by subsequent users. Although the interpretation of the resulting ICC will differ for mixed and random models (one can only generalize to subsequent raters if the raters one uses are sampled at random), the calculated value will not be affected by this distinction.

**Table 5** Analysis of Variance Summary Table for Two-Way Design

| *Source* | *df* | *Sum Sq* | *Mean Sq* | *F value* |
|---|---|---|---|---|
| Row (Couple) | 14 | 1373.233 | 98.088 | 10.225 |
| Rater | 3 | 274.850 | 91.617 | 9.551 |
| Error | 42 | 402.900 | 9.593 | |
| Total | 59 | 2050.983 | | |

#### Agreement Versus Consistency

From a computational perspective, the researcher needs to distinguish between a measure of consistency and a measure of agreement. Consistency measures consider only whether raters are aligned in their ordering of targets, whereas agreement measures take into account between-rater variance in the mean level of their ratings. If one trains graduate assistants as raters so that they can divide up the data and have each rater rate 10 different couples, it is necessary to ensure that they would give the same behavioral sample similar ratings. A measure of agreement is needed. If, instead, rater means will be equated before using the data, a measure of consistency might be enough. Agreement and consistency lead to different ICC coefficients.

#### Unit of Reliability

The final distinction is between whether the concern is the reliability of individual raters or of a mean rating for each target. It should be evident that mean ratings will be more reliable than individual ratings, and the ICC coefficient should reflect that difference.

### *Two-Way Model for Consistency*

Although the distinction between mixed and random models has interpretive importance, there will be no difference in the computation of the ICCs. For that reason, the two models will not be distinguished in what follows. But distinctions must be made on the basis of consistency versus agreement and on the unit of reliability.

Assume that the data in Table 4 represent the ratings of 15 couples (rows) by four randomly selected raters. Of concern is whether raters can use a new measurement scale to rank couples in the same order. First, assume that the ultimate unit of measurement will be the individual rating.

The analysis of variance for the data in Table 4 is shown in Table 5.

With either a random or mixed-effects model, the reliability of ratings in a two-way model for consistency is defined as

$$\mathrm{ICC}_{C,1} = \frac{MS_{\mathrm{row}} - MS_{\mathrm{error}}}{MS_{\mathrm{row}} + (k-1)MS_{\mathrm{error}}}.$$

The notation $\mathrm{ICC}_{C,1}$ refers to the ICC for consistency based on the individual rating. For the example, in Table 2 this becomes

$$\begin{aligned}\mathrm{ICC}_{C,1} &= \frac{MS_{\mathrm{row}} - MS_{\mathrm{error}}}{MS_{\mathrm{row}} + (k-1)MS_{\mathrm{error}}} \\ &= \frac{98.088 - 9.593}{98.088 + (4-1)\times 9.593} = \frac{88.495}{126.867} \\ &= .698.\end{aligned}$$

A significance test for this coefficient against the null hypothesis $\rho = 0$ is given by the test on rows (couples) in the summary table. This coefficient is clearly significantly different from 0.

$F_{\mathrm{L}}$ and $F_U$ are defined as before:

$$F_L = F_{\mathrm{obs}}/F_{(.975,14,42)} = 10.225/2.196 = 4.654.$$

$$\begin{aligned}F_U &= F_{\mathrm{obs}}(.975, 42, 14) \\ &= 10.225 \times 2.668 = 27.280.\end{aligned}$$

The 95% confidence intervals (CIs) are given by

$$\mathrm{CI}_L = \frac{F_L - 1}{F_L + (k-1)} = \frac{4.656 - 1}{4.656 + 3} = .478,$$

and

$$\mathrm{CI}_U = \frac{F_U - 1}{F_U + (k-1)} = \frac{27.280 - 1}{27.280 + 3} = .869.$$

Notice that the degrees of freedom are again reversed in computing $F_U$.

But suppose that the intent is to average the $k = 4$ ratings across raters. This requires the reliability of that average rating. Then, define

$$\mathrm{ICC}_{c,4} = \frac{MS_{\mathrm{row}} - MS_{\mathrm{error}}}{MS_{\mathrm{row}}},$$

which is .902.

The $F$ for $\rho = 0$ is still the $F$ for rows, with $F_L$ and $F_U$ defined as for the one-way. The confidence limits on $\rho$ become

$$\text{CI}_L = 1 - \frac{1}{F_L} = 1 - \frac{1}{4.656} = .785,$$

and

$$\text{CI}_U = 1 - \frac{1}{F_U} = 1 - \frac{1}{27.280} = .963.$$

It is important to think here about the implications of a measure of consistency. In this situation, $\text{ICC}_{(C,1)}$ is reasonably high. It would be unchanged (at .698) if a rater E was created by subtracting 15 points from rater D and then substituting rater E for rater D. If one were to take any of these judges (or perhaps other judges chosen at random) to a high school diving competition, their rankings should more or less agree. Then, the winners would be the same regardless of the judge. But suppose that one of these judges, either rater D or rater E, was sent to that same high school but asked to rate the reading ability of each student and make a judgment of whether that school met state standards in reading. Even though each judge would have roughly the same ranking of children, using rater E instead of rater D could make a major difference in whether the school met state standards for reading. Consistency is not enough in this case, whereas it would have been enough in diving.

### *Two-Way Models for Agreement*

The previous section was concerned not with the absolute level of ratings but only with their relative ratings. Now, consider four graduate assistants who have been thoroughly trained on using a particular rating system. They will later divide up the participants to be rated. It is important that they have a high level of agreement so that the rating of a specific couple would be the same regardless of which assistant conducted the rating.

When the concern is with agreement using a two-way model,

$$\text{ICC}_{A,1} = \frac{MS_{\text{row}} - MS_{\text{error}}}{MS_{\text{row}} + (k-1)MS_{\text{error}} + \frac{k(MS_{\text{rater}} - MS_{\text{error}})}{n}}.$$

The notation $\text{ICC}_{A,1}$ represents a measure of agreement for the reliability of individual ratings. With consistency, individual raters could use different anchor points, and rater differences were not involved in the computation. With agreement, rater differences matter, which is why $\text{MS}_{\text{rater}}$ appears in the denominator for $\text{ICC}_{A,1}$.

Using the results presented in Table 5 gives

$$\begin{aligned}\text{ICC}_{A,1} &= \frac{MS_{\text{row}} - MS_{\text{error}}}{MS_{\text{row}} + (k-1)MS_{\text{error}} + \frac{k(MS_{\text{rater}} - MS_{\text{error}})}{n}}, \\ &= \frac{98.088 - 9.593}{98.088 + (4-1)9.593 + \frac{4(91.617 - 9.593)}{15}}, \\ &= .595.\end{aligned}$$

The test on $H_0 : \rho = 0$ is given by the $F$ for rows in the summary table and is again 10.255, which is significant on $(n-1)$ and $k(n-1)$ *df*. However the calculation of confidence limits in this situation is complex, and the reader is referred to McGraw and Wong (1996) for the formulas involved.

If interest is, instead, the reliability of mean ratings, then

$$\begin{aligned}\text{ICC}_{A,4} &= \frac{MS_{\text{row}} - MS_{\text{error}}}{MS_{\text{row}} + \frac{(MS_{\text{rater}} - MS_{\text{error}})}{n}}, \\ &= \frac{98.088 - 9.593}{98.088 + \frac{(91.617 - 9.593)}{15}} = .855.\end{aligned}$$

An $F$ test on the statistical significance of $\text{ICC}_{A,k}$ is given in Table 8 of McGraw and Wong, and corresponding confidence limits on the estimate are given in Table 7 of McGraw and Wong.

In the discussion of consistency creating a rater E by subtracting 15 points from rater D and then replacing rater D with rater E led to no change in $\text{ICC}_{C,1}$. But with that same replacement, the agreement ICCs would be $\text{ICC}_{A,1} = .181$, and $\text{ICC}_{A,4} = .469$. Clearly, the agreement measure is sensitive to the mean rating of each rater.

*David C. Howell*

***See also*** Classical Test Theory; Coefficient Alpha; Correlation; Mixed Model Design; Reliability

**Further Readings**

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1,* 30–46.

Nichols, D. (1998). *Choosing an intraclass correlation coefficient.* Retrieved June 16, 2009, from http://www.ats.ucla.edu/stat/spss/library/whichicc.htm

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing reliability. *Psychological Bulletin, 86,* 420–428.

# Item Analysis

Item analysis is the set of qualitative and quantitative techniques and procedures used to evaluate the characteristics of items of the test before and after the test development and construction. An item is a basic building block of a test, and its analysis provides information about its performance. Item analysis allows selecting or omitting items from the test, but more important, item analysis is a tool to help the item writer improve an item. Anthony Nitko suggests that some classroom uses of an item analysis might be estimating whether an item functions as intended, providing feedback to students about their performance, providing feedback to the teacher about student difficulties and ideas of curriculum improvement, revising assessment tasks, and improving item-writing skills.

Item analysis can be used both for dichotomously scored (correct or incorrect) items and polytomously scored (with more than two score categories) items. The main purpose of item analysis is to improve internal consistency or internal structure validity, focused on confirming a single-factor or one-trait test. If the trait is not one factor, then the use of item analysis might tend to lower validity. If a test has two factors (or content divisions) or is multifactored (more than two content divisions), the calculation of item statistics for each item (or option) should be focused on the subtotal for the relevant set of items rather than on the total test score. Item analysis in this case is used to improve the internal consistency of each subset of items with no intention to change the dimensionality of the entire set. In these cases, an overall reliability index would be stratified alpha (or battery alpha) rather than the regular coefficient alpha as used for a one-dimensional test.

A test that is composed of items selected based on item analysis statistics tends to be more reliable than one composed of an equal number of unanalyzed items. Even though the process of item analysis looks sophisticated, it is not as challenging as it seems. Item analysis software programs are user friendly and make the task much simpler. The most time-consuming part of item analysis might be the process of tabulating the data to be analyzed by the software program. Statistical programs such as SPSS, an IBM product, and SAS can be used for item analysis as well. The information needed to start an item analysis procedure after the administration is the work of the examinees. In the subsequent sections, several aspects of the item analysis process are discussed.

## Norm-Referenced Versus Criterion-Referenced Interpretations

Norm-referenced and criterion-referenced interpretations are two different types of score interpretations. Norm-referenced interpretations help to locate an examinee's position within a well-defined group, and criterion-referenced interpretations are used in describing a degree of proficiency in a specified content domain. Use of college admission tests is an example of norm-referenced interpretations, and most tests and quizzes written by teachers are examples of criterion-referenced interpretations. A criterion might be the lesson curriculum or the state standards. Some tests are intended for both types of interpretations; for example, some states use standards-based tests for both purposes. With standards-based tests, the criterion-referenced interpretation is intended to give information about how proficient the students are in the curriculum defined by the state standards. The norm-referenced interpretation provides a measure of how each student compares with peers.

### *Norm-Referenced Interpretations*

Norm-referenced interpretations are involved with determining how an examinee's score compares with others who have taken the same test. Item analysis techniques and procedures for norm-referenced interpretations are employed to refine

and improve the test by identifying items that will increase the ability of the test to discriminate among the scores of those who take the test. In norm-referenced interpretations, the better an item discriminates among examinees within the range of interest, the more information that item provides. Discrimination among examinees is the most crucial characteristic desired in an item used for a norm-referenced purpose. The discriminating power is determined by the magnitude of the item discrimination index that will be discussed next.

### Discrimination Index

The item discrimination index, which is usually designated by the uppercase letter D (also net D, U-L, ULI, and ULD), shows the difference between upper and lower scorers answering the item correctly. It is the degree to which high scorers were inclined to get each item right and low scorers were inclined to get that item wrong. D is a measure of the relationship between the item and the total score, where the total score is used as a substitute for a criterion of success on the measure being assessed by the test. For norm-referenced interpretations, there is usually no available criterion of success because that criterion is what the test is supposed to represent.

Upper (U) and lower (L) groups can be decided by dividing the arranged descending scores into three groups: upper, middle, and lower. When there is a sufficient number of normally distributed scores, Truman Kelley demonstrated that using the top and bottom 27% (1.225 standard deviation units from the mean) of the scores as upper and lower groups would be the best to provide a wide difference between the groups and to have an adequate number of scores in each group. When the total number of scores is between 20 and 40, it is advised to select the top 10 and the bottom 10 scores. When the number of scores is less than or equal to 20, two groups are used without any middle group.

One way of computing the item discrimination index is finding the difference of percentages of correct responses of U and L groups by computing Up − Lp (U percentage minus L percentage). Typically, this percentage difference is multiplied by 100 to remove the decimals; the result is D, which yields values with the range of $+100$ and $-100$. The maximum value of D ($+100$) occurs when all examinees in the upper group get an item right and all examinees in the lower group fail. The value of D equals 0 when the item is correctly answered by all, none, or any other same percentage of examinees in both upper and lower groups. D has a negative value when the percentage of students answering the item correctly in the lower group is greater than the percentage of correct responses in the upper group.

A test with items having high D values produces more spread in scores, therefore contributing to the discrimination in ability among examinees. The item discrimination index is the main factor that directly affects item selection in norm-referenced tests. Ideally, items selected for norm-referenced interpretations are considered good with D values above 30 and very good with values above 40. The reliability of the test will be higher by selecting items that have higher item discrimination indices.

Warren Findley demonstrated that the index of item discrimination is absolutely proportional to the difference between the numbers of correct and incorrect discriminations (bits of information) of the item. Assuming 50 individuals in the upper and 50 individuals in the lower groups, the ideal item would distinguish each of 50 individuals in the U group from each in the L group. Thus, $50 \times 50 = 2,500$ possible correct discriminations or bits of information. Considering an item on which 45 individuals of the upper group but only 20 individuals of the lower group answer the item correctly, the item would distinguish 45 individuals who answered correctly from the upper group from 30 individuals who answered incorrectly in the lower group, generating a total of $45 \times 30 = ,350$ correct discriminations. Consequently, 5 individuals answering the item incorrectly in the upper group are distinguished incorrectly from the 20 individuals who answered correctly in the lower group, generating $5 \times 20 = 100$ incorrect discriminations. The net amount of effective discriminations of the item is $1,350 - 100 = 1,250$ discriminations, which is 50% of the 2,500 total maximum possible correct discriminations. Note that the difference between the number of examinees who answer the item correctly in the upper group and in the lower group is $45 - 20 = 25$, which is half of the maximum possible difference with 50 in each group.

Another statistic used to estimate the item discrimination is the point-biserial correlation coefficient. The point-biserial correlation is obtained when the Pearson product-moment correlation is computed from a set of paired values where one of the variables is dichotomous and the other is continuous. Correlation of passing or failing an individual item with the overall test scores is the common example for the point-biserial correlation in item analysis. The Pearson correlation can be calculated using SPSS or SAS. One advantage cited for using the point-biserial as a measure of the relationship between the item and the total score is that it uses all the scores in the data rather than just the scores of selected groups of students.

#### Difficulty Index

The item difficulty index, which is denoted as $p$, is the proportion of the number of examinees answering the item correctly to the total number of examinees. The item difficulty index ranges from 0 to 1. Item difficulty also can be presented as a whole number by multiplying the resulting decimal by 100. The difficulty index shows the percentage of examinees who answer the item correctly, although the easier item will have a greater value. Because the difficulty index increases as the difficulty of an item decreases, some have suggested that it be called the easiness index. Robert Ebel suggested computing item difficulty by finding the proportion of examinees that answered the item incorrectly to the total number of examinees. In that case, the lower percentage would mean an easier item.

It is sometimes desirable to select items with a moderate spread of difficulty and with an average difficulty index near 50. The difficulty index of 50 means that half the examinees answer an item correctly and other half answer incorrectly. However, one needs to consider the purpose of the test when selecting the appropriate difficulty level of items. The desired average item difficulty index might be increased with multiple-choice items, where an examinee has a chance to guess. The item difficulty index is useful when arranging the items in the test. Usually, items are arranged from the easiest to the most difficult for the benefit of examinees.

In norm-referenced interpretations, an item does not contribute any information regarding the difference in abilities if it has a difficulty index of 0, which means none of the examinees answer correctly, or 100, which means all of the examinees answer correctly. Thus, items selected for norm-referenced interpretations usually do not have difficulty indices near the extreme values of 0 or 100.

### *Criterion-Referenced Interpretations*

Criterion-referenced interpretations help to interpret the scores in terms of specified performance standards. In psychometric terms, these interpretations are concerned with absolute rather than relative measurement. The term *absolute* is used to indicate an interest in assessing whether a student has a certain performance level, whereas *relative* indicates how a student compares with other students. For criterion-referenced purposes, there is little interest in a student's relative standing within a group.

Item statistics such as item discrimination and item difficulty as defined previously are not used in the same way for criterion-referenced interpretations. Although validity is an important consideration in all test construction, the content validity of the items in tests used for criterion-referenced interpretations is essential. Because in many instances of criterion-referenced testing the examinees are expected to succeed, the bell curve is generally negatively skewed. However, the test results vary greatly depending on the amount of instruction the examinees have had on the content being tested. Item analysis must focus on group differences and might be more helpful in identifying problems with instruction and learning rather than guiding the item selection process.

#### Discrimination Index

A discrimination index for criterion-referenced interpretations is usually based on a different criterion than the total test score. Rather than using an item–total score relationship as in norm-referenced analysis, the item–criterion relationship is more relevant for a criterion-referenced analysis. Thus, the upper and lower groups for a discrimination index should be selected based on their performance on the criterion for the standards or curriculum of interest. A group of students who have mastered the skills of interest could comprise the upper group, whereas those who have not yet learned

those skills could be in the lower group. Or a group of instructed students could comprise the upper group, with those who have had no instruction on the content of interest comprising the lower group. In either of these examples, the D index would represent a useful measure to help discriminate the masters versus the nonmasters of the topic.

After adequate instruction, a test on specified content might result in very little score variation with many high scores. All of the examinees might even get perfect scores, which will result in a discrimination index of zero. But all these students would be in the upper group and not negatively impact the discrimination index if calculated properly.

Similarly, before instruction, a group might all score very low, again with very little variation in the scores. These students would be the lower group in the discrimination index calculation. Within each group, there might be no variation at all, but between the groups, there will be evidence of the relationship between instruction and success. Thus, item discrimination can be useful in showing which items measure the relevance to instruction, but only if the correct index is used.

As opposed to norm-referenced interpretations, a large variance in performance of an instructed group in criterion-referenced interpretation would probably indicate an instructional flaw or a learning problem on the content being tested by the item. But variance between the instructed group and the not-instructed group is desired to demonstrate the relevance of the instruction and the sensitivity to the test for detecting instructional success.

### Difficulty Indexes

The difficulty index of the item in criterion-referenced interpretations might be a relatively more useful statistic than discrimination index in identifying what concepts were difficult to master by students. In criterion-referenced testing, items should probably have an average difficulty index around 80 or 90 within instructed groups. It is important to consider the level of instruction a group has had before interpreting the difficulty index.

In criterion-referenced interpretations, a difficulty index of 0 or 100 is not as meaningless as in norm-referenced interpretations, because it must be known for what group the index was obtained. The difficulty index of zero would mean that none of the examinees answered the item correctly, which is informative regarding the "nonmastered" content and could be expected for the group that was not instructed on this content. The difficulty index of 100 would mean that all the examinees answered an item correctly, which would confirm the mastery of content by examinees who had been well taught. Items with difficulties of 0 or 100 would be rejected for norm-referenced purposes because they do not contribute any information about the examinee's relative standing. Thus, the purpose of the test is important in using item indices.

## Effectiveness of Distractors

One common procedure during item analysis is determining the performance of the distractors (incorrect options) in multiple-choice items. Distractors are expected to enhance the measurement properties of the item by being acceptable options for the examinees with incomplete knowledge of the content assessed by the item. The discrimination index is desired to be negative for distractors and intended to be positive for the correct options. Distractors that are positively correlated with the test total are jeopardizing the reliability of the test; therefore, they should be replaced by more appropriate ones. Also, there is percent marked, (percentage upper + percentage lower)/2, for each option as a measure of "attractiveness." If too few examinees select an option, its inclusion in the test might not be contributing to good measurement unless the item is the one indicating mastery of the subject. Also, if a distractor is relatively more preferred among the upper group examinees, there might be two possible correct answers.

A successful distractor is the one that is attractive to the members of the low-scoring group and not attractive to the members of the high-scoring group. When constructing a distractor, one should try to find the misconceptions related to the concept being tested. Some methods of obtaining acceptable options might be use of context terminology, use of true statements for different arguments, and inclusion of options of similar difficulty and complexity.

### Differential Item Functioning

The difficulty index should not be confused with differential item functioning (DIF). DIF analysis investigates every item in a test for the signs of interactions with sample characteristics. Differential item functioning occurs when people from the different groups (race or gender) with the same ability have different probabilities of giving a correct response on an item. An item displays DIF when the item parameters such as estimated item difficulty or item discrimination index differ across the groups.

Because all distractors are incorrect options, the difference among the groups in distractor choice has no effect on the test score. However, group difference in the distractor choice might indicate that the item functions differently for the different subgroups. Analysis of differential distractor functioning (DDF) examines only incorrect responses. Although DIF and DDF are not usually considered under item analysis, they are essential investigations for examining validity of tests.

### Rasch Model

The item analysis considered to this point has been focused on what is called classical test theory. Another approach to test development is called item response theory (IRT). There are several models of IRT that are beyond the scope of this writing. The one-parameter model, called the Rasch model, deserves mention as an introduction to IRT.

Item analysis in the Rasch model is based on the item characteristic curve (ICC) of the one-parameter logistic model. The ICC is a nonlinear regression of the probability of correct response for the dichotomous item (with the range of 0 to 1) on the ability (trait or skill) to be measured (with the range of $-\infty$ to $+\infty$). The placement of the curve is related to the difficulty index, and the slope of the curve is related to the D index discussed previously. According to the Rasch model, items having same discrimination indices but different difficulty indices should be selected for the test. The Rasch model is usually used for developing tests intended for norm-referenced interpretations or for tests like standards-based tests that are intended for both types of interpretations.

*Perman Gochyyev and Darrell Sabers*

*See also* Differential Item Functioning; Internal Consistency Reliability; Item Response Theory; Item-Test Correlation; Pearson Product-Moment Correlation Coefficient; Reliability; SAS; SPSS

#### Further Readings

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Nitko, A. J. (2004). *Educational assessment of students* (4th ed.). Upper Saddle River, NJ: Pearson.

## Item Response Theory

Item response theory (IRT) is a mental measurement theory based on the postulate that an individual's response to a test item is a probabilistic function of characteristics of the person and characteristics of the item. The person characteristics are the individual's levels of the traits being measured, and the item characteristics are features such as difficulty and discriminating power. Item response theory has several advantages over classic test theory and has the potential to solve several difficult measurement problems. The foundations of item response theory were developed in the early 20th century; however, it was Frederic Lord, beginning in the 1950s, who organized and developed the theory into a framework that could be applied to practical testing problems. Advances in computing were necessary to make the theory accessible to researchers and practitioners. Item response theory is now widely used in educational contexts by testing companies, public school systems, the military, and certification and licensure boards, and is becoming more widely used in other contexts such as psychological measurement and medicine. This entry discusses item response models and their characteristics, estimation of parameters and goodness of fit of the models, and testing applications.

### Item Response Models

Item response theory encompasses a wide range of models depending on the nature of the item score,

the number of dimensions assumed to underlie performance, the number of item characteristics assumed to influence responses, and the mathematical form of the model relating the person and item characteristics to the observed response. The item score might be dichotomous (correct/incorrect), polytomous as in multiple-choice response or graded performance scoring, or continuous as in a measured response. Dichotomous models have been the most widely used models in educational contexts because of their suitability for multiple choice tests. Polytomous models are becoming more established as performance assessment becomes more common in education. Polytomous and continuous response models are appropriate for personality or affective measurement. Continuous response models are not well known and are not discussed here.

The models that are currently used most widely assume that there is a single trait or dimension underlying performance; these are referred to as unidimensional models. Multidimensional models, although well-developed theoretically, have not been widely applied. Whereas the underlying dimension is often referred to as "ability," there is no assumption that the characteristic is inherent or unchangeable.

Models for dichotomous responses incorporate one, two, or three parameters related to item characteristics. The simplest model, which is the one-parameter model, is based on the assumption that the only item characteristic influencing an individual's response is the difficulty of the item. A model known as the Rasch model has the same form as the one-parameter model but is based on different measurement principles. The Rasch theory of measurement was popularized in the United States by Benjamin Wright. The two-parameter model adds a parameter for item discrimination, reflecting the extent to which the item discriminates among individuals with differing levels of the trait. The three-parameter model adds a lower asymptote or pseudo-guessing parameter, which gives the probability of a correct response for an individual with an infinitely low level of the trait.

The earliest IRT models used a normal ogive function to relate the probability of a correct response to the person and item characteristics. Although the normal ogive model is intuitively appealing and provides a connection between IRT item parameters and classic item indices, an integral must be calculated to obtain the probability of a correct response. Allen Birnbaum proposed a more mathematically tractable cumulative logistic function. With an appropriate scaling factor, the normal ogive and logistic functions differ by less than .05 over the entire trait continuum. The logistic model has become widely accepted as the basic item response model for dichotomous and polytomous responses.

The unidimensional three-parameter logistic model for dichotomous responses is given by

$$P(u_j = 1|\theta) \;=\; c_j + (1 - c_j)\frac{e^{1.7a_j(\theta - b_j)}}{1 + e^{1.7a_j(\theta - b_j)}},$$

where $u_j$ is the individual's response to item $j$, scored 1 for correct and 0 for incorrect, $\theta$ is the individual's value on the trait being measured, $P(u_j = 1|\theta)$ is the probability of a correct response to item $j$ given $\theta$, $c_j$ is the lower asymptote parameter, $a_j$ is the item discrimination parameter, $b_j$ is the item difficulty parameter, and 1.7 is the scaling factor required to scale the logistic function to the normal ogive. The curve produced by the model is called the item characteristic curve (ICC). ICCs for several items with differing values of the item parameters are shown in Figure 1. The two-parameter model is obtained by omitting the lower asymptote parameter, and the one-parameter model is obtained by subsequently omitting the discrimination parameter. The two-parameter model assumes that low-performing individuals have no chance of answering the item correctly through guessing, whereas the one-parameter model assumes that all items are equally discriminating.

The lower asymptote parameter is bounded by 0 and 1 and is usually less than .3 in practice. The discrimination parameter is proportional to the slope of the curve at its point of inflection; the steeper the slope, the greater the difference in probability of correct response for individuals of different trait levels, hence, the more discriminating the item. Discrimination parameters must be positive for valid measurement. Under the one- and two-parameter models, the difficulty parameter represents the point on the trait continuum where the probability of a correct response is 0.5;

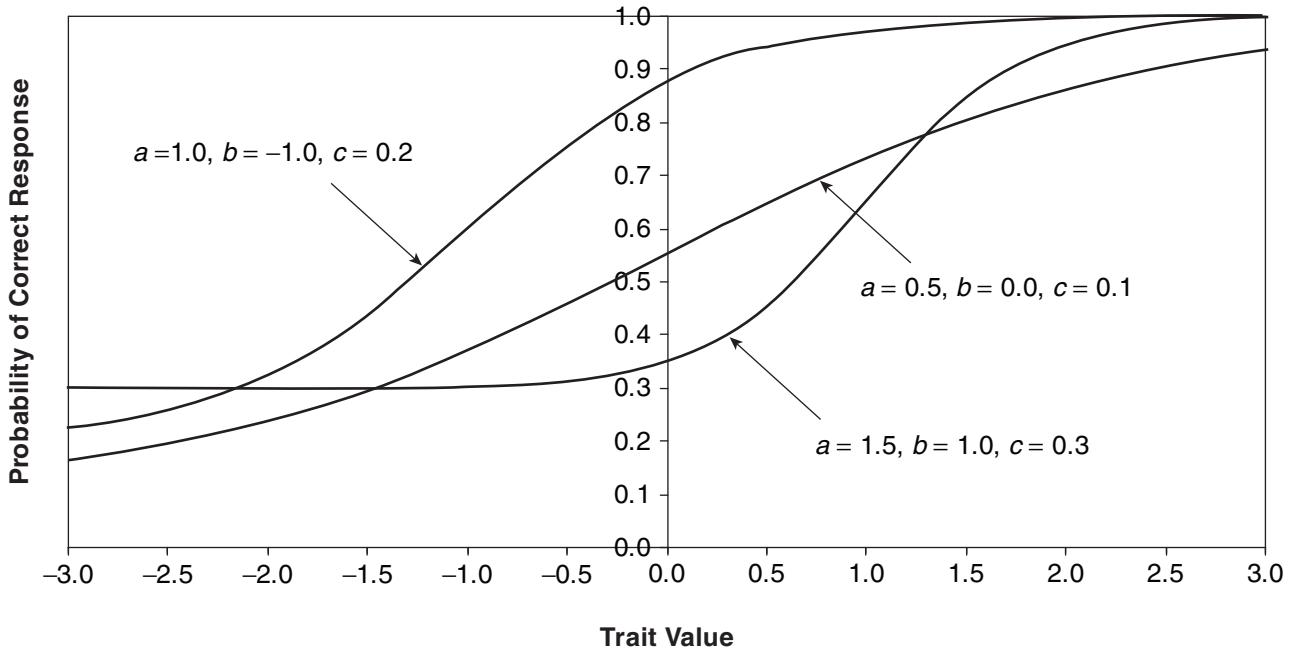



**Figure 1** Item Characteristic Curves for Three Items Under the Three-Parameter Logistic Item Response Theory Model

under the three-parameter model, the probability of a correct response at $\theta = b_j$ is $(1 + c_j)/2$.

Note the indeterminacy in the model previously: The model does not specify a scale for $a, b,$ and $\theta$. A linear transformation of parameters will produce the same probability of a correct response, that is, if $\theta^* = A\theta + B$, $b^* = Ab + B$, and $a^* = a/A$, then $a^*(\theta^* - b^*) = a(\theta - b)$ and $P(\theta^*) = P(\theta)$. The scale for parameter estimates is typically fixed by standardizing on either the $\theta$ values or the $b$ values. With this scaling, the $\theta$ and $b$ parameter estimates generally fall in the range $(-3, 3)$ and the $a$ parameter estimates are generally between 0 and 2.

There are several item response models for polytomous responses. When there is no assumption that the response categories are on an ordered scale, the nominal response model might be used to model the probability that an individual will score in a particular response category. The nominal response model is not widely used because polytomously scored responses are generally ordered in practice, as, for example, in essay or partial credit scoring. There are several models for ordered polytomous responses: The most well known are the graded response model (GRM), the partial credit model (PCM), the generalized partial credit model (GPCM), and the rating scale model (RSM). Only the GRM and the GPCM are described here.

The GRM is obtained by formulating two-parameter dichotomous models for the probability that an examinee will score in each response category or higher (as opposed to a lower category), then subtracting probabilities to obtain the probability of scoring within each response category, that is,

$$P(u_j \geq k|\theta) = P_k^*(\theta) = \frac{e^{a_j(\theta - b_{jk})}}{1 + e^{a_j(\theta - b_{jk})}},$$
$$k = 1, \ldots, m-1, P_0^*(\theta) = 1,$$
$$P_m^*(\theta) = 0,$$
$$P(u_j = k) = P_k^*(\theta) - P_{k+1}^*(\theta).$$

Here, responses are scored 0 through $m - 1$, where $m$ is the number of response categories, $k$ is the response category of interest, $a_j$ is the discrimination parameter, interpreted as in dichotomous models, and $b_{jk}$ is the category parameter. The

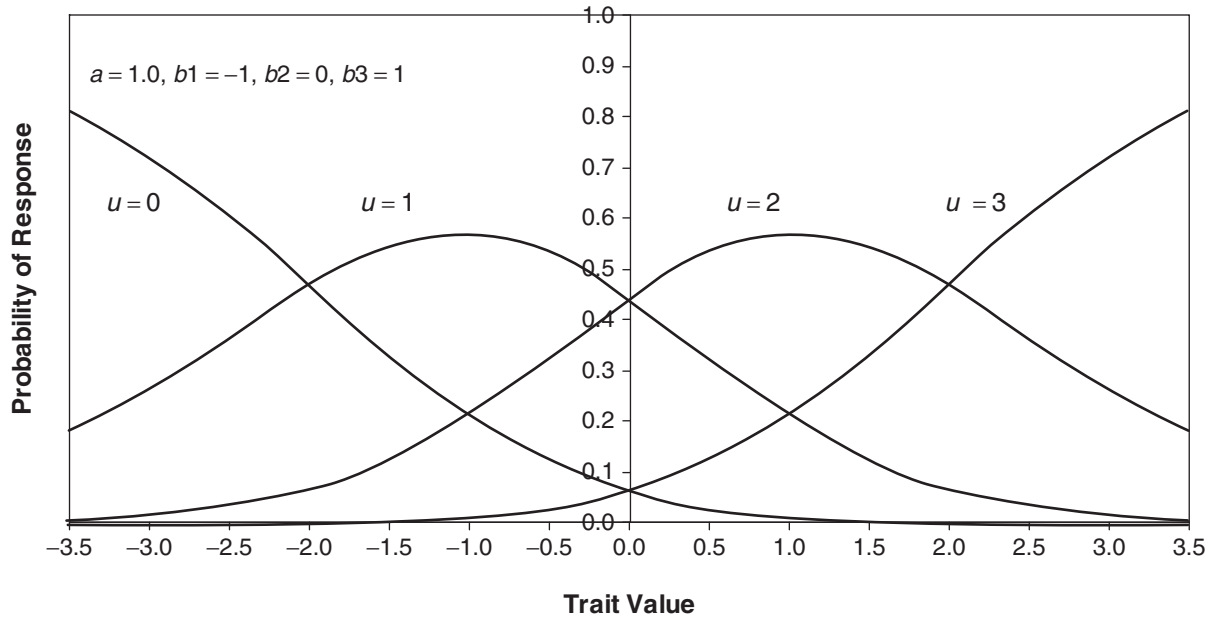



**Figure 2** Item Response Category Characteristic Curves for an Item Under the GPCM

category parameters represent the level of the trait needed to have a 50% chance of scoring in that category or higher. Category parameters are necessarily ordered on the trait scale. It is assumed that the item is equally discriminating across category boundaries. The model provides a separate item response function for each response category; the resultant curves are called item response category characteristic curves (IRCCs).

The GPCM differs from the GRM in that it is based on a comparison of adjacent categories. The model is given by

$$P(u_j = k|\theta) = P_{jk}(\theta), = \frac{e^{a_j \sum_{v=1}^{k} (\theta - b_{jv})}}{1 + \sum_{c=1}^{m-1} e^{a_j \sum_{v=1}^{c} (\theta - b_{jv})}},$$

$$k = 1, \ldots, m-1; P_{j0}(\theta) = \frac{1}{1 + \sum_{c=1}^{m-1} e^{a_j \sum_{v=1}^{c} (\theta - b_{jv})}}.$$

In this model, the category parameter $b_{jk}$ represents the trait value at which an individual has an equal probability of scoring in category $k$ versus category $(k - 1)$. The category parameters need not be ordered under this model. The GPCM is a generalization of the PCM, which assumes equal discriminations across items and omits the discrimination parameter in the model. An example of IRCCs for a polytomous item under the GPCM is shown in Figure 2.

Item response theory has several advantages over classic test theory in measurement applications. First, IRT item parameters are invariant across subpopulations, whereas classic item indices change with the performance level and heterogeneity of the group taking the test. Person parameters are invariant across subsets of test items measuring the same dimension; whether the test is easy or hard, an individual's trait value remains the same. This is not the case with total test score, which depends on the difficulty of the test. The invariance property is the most powerful feature of item response models and provides a solid theoretical base for applications such as test construction, equating, and adaptive testing. Note that invariance is a property of the *parameters* and holds only in the *population; estimates* will vary across *samples* of persons or items.

A second advantage of IRT is individualized standard errors of measurement, rather than a group-based measure such as is calculated in

classic test theory. Another advantage is that item response theory is formulated at the item level and gives a basis for prediction of an individual's or group's performance when presented with new test items; this is useful in constructing tests for specific populations or purposes. Classic test theory is based on total test score and offers no basis for prediction of performance.

## Estimation of Parameters

Estimation of the parameters of an item response model is a computer-intensive task generally requiring large samples. When item parameters are known and only person parameters must be estimated, or vice versa, the procedure is relatively simple via maximum likelihood estimation. Maximum likelihood estimation involves computing the likelihood of the observed data as a function of the unknown parameters, based on the model to be fitted, and then determining the parameter values that maximize the likelihood.

When both item parameters and person parameters are unknown, estimation is considerably more difficult. Large samples (on the order of a thousand or more) are required to obtain adequate estimates under the three-parameter model. The most widely used commercially available computer programs for IRT parameter estimation (BILOG-MG, PARSCALE, and MULTILOG) employ marginal maximum likelihood procedures with the Expectation-Maximization (EM) algorithm. Under marginal maximum likelihood estimation, item parameters are estimated assuming the population distribution of person parameters is known. The EM algorithm is used to solve the marginal likelihood equations obtained for the item parameters. These equations involve quantities that are unknown from the data; the expected values of these quantities are substituted (E-step) and maximum likelihood estimation (M-step) is then performed. Prior distributions are often specified for the item parameters to facilitate estimation. Person parameters are subsequently estimated holding the item parameters fixed at their estimated values. A normal prior distribution is specified for $\theta$ and the mean of the posterior distribution for each individual is taken as the estimate of $\theta$. This estimate is referred to as the *expectation a posteriori* (EAP) estimate.

## Goodness of Fit

Item response theory is based on strong assumptions: that the assumed dimensionality is correct and that the mathematical model correctly specifies the relationship between the item and person parameters and response to the item. The advantages of IRT are obtained only when these assumptions are met, that is, when the model fits the data. Applications of item response theory must begin with an assessment of the fit of the model. Assessment of goodness of fit requires checking the assumptions of the model, the expected features of the model, and predictions based on the model.

The primary assumption of the models in common use is that of unidimensionality. A variety of methods has been proposed for assessing the dimensionality of a mental measurement scale. Linear factor analysis methods are most commonly used, although these fail to take into account the nonlinearity of the relationship between the trait and the observed response. Nonlinear factor analysis procedures are more appropriate but less widely used. The computer programs NOHARM and TESTFACT fit multidimensional nonlinear factor models to item response data. TESTFACT provides a chi-square test of fit of the model. William Stout developed a test of "essential unidimensionality" based on the principle of local independence, which states that after conditioning on all traits underlying performance, item responses are statistically independent of each other.

The primary expected feature of an IRT model is parameter invariance. Checking for item parameter invariance is done by comparing item parameter estimates obtained in different subgroups of examinees; estimates should differ by no more than sampling error. Likewise, ability estimates based on different subsets of items should also be similar.

Checking predictions based on the model can be done at the test level or item level. At the test level, observed score distributions can be compared with predicted score distributions based on the fitted model. At the item level, differences between observed and expected proportions of examinees in each response category within subgroups based on trait estimates or expected score can be examined graphically or by means of chi-square fit statistics.

## Applications

Item response theory provides a natural framework for many testing applications, including test construction. Test and item information are important concepts for this purpose. The information provided by the test about an individual with a given value of $\theta$ is inversely related to the standard error of estimate of $\theta$, that is,

$$I(\theta) = \frac{1}{\sqrt{SE(\theta)}}.$$

Test information is a sum over items; hence, each item contributes independently to the information provided by the test. This additive property allows the selection of items to create a test that measures with a desired degree of precision in any part of the trait continuum. Items provide their maximum information at a trait value equal to the difficulty of the item. More discriminating items provide more information, and for dichotomous items, lower $c$ parameters increase information.

Equating of test forms is another area in which IRT provides an elegant solution. Because person parameters are invariant across subsets of items, the scores of individuals who have taken different forms of a test can be validly compared. However, because of the indeterminacy problem, it is necessary to ensure that the estimates are on a common scale. This step is most often accomplished by including a set of common items in the two forms to be equated, and determining the linear transformation that would equalize the means and standard deviations of the common $b$ values or equalize the test characteristic curves. This transformation is applied to the $b$ values and a corresponding transformation to the $a$ values of all items in the form to be equated to the base form. The same transformation applied to the $b$ values is applied to the trait estimates of the group taking the form to be equated to place them on the same scale as that in the reference group.

Another area in which IRT is useful is in the detection of differential item functioning (DIF) for ethnic, gender, and other demographic subgroups. An item shows differential functioning if individuals at the same trait value do not have the same probability of scoring in a given response category. When this occurs, it indicates that group membership is an additional dimension influencing performance, a violation of model assumptions. In this case, item parameter invariance is not obtained across subgroups and the ICCs for the subgroups differ. Differences between ICCs can be quantified by calculating the area between the curves or by computing a chi-square statistic for testing the equality of parameters.

The invariance property of IRT also provides a basis for adaptive testing. In adaptive testing, individuals are administered items one by one from a large precalibrated, equated pool and their trait estimates are updated after each item based on their response. Each subsequent item is selected to provide maximum information at the individual's current estimated trait value. Testing can be terminated when the standard error of the $\theta$ estimate falls below a preset criterion. Although each person has taken a different test, trait estimates are comparable. Adaptive testing has several advantages, among which are equal measurement precision for all examinees, shorter tests, immediate score reporting, and potentially greater test security.

Item response theory provides a powerful framework for mental measurement. When the assumptions of the model are met, IRT can provide elegant solutions to many measurement problems.

*H. Jane Rogers*

***See also*** Classical Test Theory; Computerized Adaptive Testing; Differential Item Functioning; Psychometrics

### Further Readings

Andrich D. (1988). *Rasch models for measurement.* Newbury Park, CA: Sage.

Baker, F. B., & Kim, S. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York: Marcel-Dekker.

Du Toit, M. (2003). *IRT from SSI: BILOG MG, MULTILOG,PARSCALE, TESTFACT.* Lincolnwood, IL: Scientific Software International.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum.

Hambleton, R. K., & Swaminathan H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer-Nijohoff.

Hambleton R. K., Swaminathan, H., & Rogers H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.
Hays, R. D., Morales, L. S., & Reise, S. P. (2000). Item response theory and health outcomes measurement in the 21st century. *Medical Care, 38,* II28–II42.
Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement.* Homewood, IL: Dow Jones-Irwin.
Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer-Verlag.
Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum.
Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.
Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models.* Thousand Oaks, CA: Sage.
Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Chicago: University of Chicago Press.
Van der Linden, W. J., & Glas, C. A.W. (2000). *Computerized adaptive testing: Theory and practice.* Boston: Kluwer.
Wainer, H. (Ed.). (2000). *Computerized adaptive testing: A primer* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.
Wright B. D., & Stone M. H. (1979). *Best test design: Rasch measurement.* Chicago: MESA Press.

## Item-Test Correlation

The item-test correlation is the Pearson correlation coefficient calculated for pairs of scores where one item of each pair is an item score and the other item is the total test score. The greater the value of the coefficient, the stronger is the correlation between the item and the total test. Test developers strive to select items for a test that have a high correlation with the total score to ensure that the test is internally consistent. Because the item-test correlation is often used to support the contention that the item is a "good" contributor to what the test measures, it has sometimes been called an *index of item validity.* That term applies only to a type of evidence called *internal structure validity,* which is synonymous with *internal consistency reliability.* Because the item-test correlation is clearly an index of internal consistency, it should be considered as a measure of item functioning associated with that type of reliability. The item-test correlation is one of many item discrimination indices used in item analysis.

Because item responses are typically scored as zero when incorrect and unity (one) if correct, the item variable is binary or dichotomous (having two values). The resulting correlation is properly called a point-biserial coefficient when a binary item is correlated with a total score that has more than two values (called polytomous or continuous). However, some items, especially essay items, performance assessments, or those for inclusion in affective scales, are not usually dichotomous, and thus some item-test correlations are regular Pearson coefficients between polytomous items and total scores. The magnitude of correlations found when using polytomous items is usually greater than that observed for dichotomous items. Reliability is related to the magnitude of the correlations and to the number of items in a test, and thus with polytomous items, a lesser number of items is usually sufficient to produce a given level of reliability. Similarly, to the extent that the average of the item-test correlations for a set of items is increased, the number of items needed for a reliable test is reduced.

All correlations tend to be higher in groups that have a wide range of talent than in groups where there is a more restricted range. In that respect, the item-test correlation presents information about the group as well as about the item and the test. The range of talent in the group might be limited in some samples, for example, in a group of students who have all passed prerequisites for an advanced class. In groups where a restriction of range exists, the item-test correlations will provide a lower estimate of the relationship between the item and the test.

When the range of talent in the group being tested is not restricted, the item-test correlation is a spurious measure of item quality. The spuriousness arises from the inclusion of the particular item in the total test score, resulting in the correlation between an item and itself being added to the correlation between the item and the rest of the total test score. A preferred concept might be the *item–rest correlation,* which is the correlation between the item and the sum of the rest of the item scores. Another term for this item-rest correlation is the

*corrected item-test correlation,* the name given to this type of index in the SPSS Scale Reliability analysis (SPSS, an IBM company).

The intended use of the test is an important factor in interpreting the magnitude of an item-test correlation. If the intention is to develop a test with high criterion-related test validity, one might seek items that have high correlations with the external criterion but relatively lower correlations with the total score. Such items presumably measure aspects of the criterion that are not adequately covered by the rest of the test and could be preferred to items correlating highly with both the criterion and the test score. However, unless item-test correlations are substantial, the tests composed of items with high item-criterion correlations might be too heterogeneous in content to provide meaningful interpretations of the test scores. Thus, the use of item-test correlations to select items for a test is based on the goal to establish internal consistency reliability rather than direct improvement in criterion validity.

The item-test correlation resembles the loading of the item on the first principal component or the unrotated first component of an analysis of all the items in a test. The concepts are related but the results are not identical in these different approaches to representing the "loading" or "impact" of an item. However, principal components analysis might help the reader to understand the rationale of measuring the relationship between an item and some related measure. Typically, using any of these methods, the researcher wants the item to represent the same trait as the total test, component, or factor of interest. This description is limited to a single-factor test or a single-component measure just as the typical internal consistency reliability is reported for a homogeneous test measuring a single trait. If the trait being measured is not a unitary trait, then other approaches are suggested because the regular item–total correlation will underestimate the value of an item.

If a researcher has a variable that is expected to measure more than one trait, then the item–total correlation can be obtained separately for each subset of items. In this approach, each total represents the subtotal score for items of a subset rather than a total test score. Using item–subtotal correlations, the researcher might develop (or examine) a measure that was internally consistent for each subset of items. To get an overall measure of reliability for such a multitrait test, a stratified alpha coefficient would be the desired reliability estimate rather than the regular coefficient alpha as reported by SPSS Scale Reliability. In the case of multitrait tests, using a regular item–total correlation for item analysis would likely present a problem because the subset with the most items could contribute too much to the total test score. This heavier concentration of similar items in the total score would result in the items in the other subsets having lower item-test correlations and might lead to their rejection.

In analyzing multiple-choice items to determine how each option might contribute to the reliability of the total test, one can adapt the item-test correlation to become an option-test correlation. In this approach, each option is examined to determine how it correlates with the total test (or subtest). If an option is expected to be a distractor (wrong response), the option–test correlation should be negative (this assumes the option is scored 1 if selected, 0 if not selected). Distractors that are positively correlated with the test or subtest total are detracting from the reliability of the measure; these options should probably be revised or eliminated.

The item-test correlation was first associated with the test analysis based on classic test theory. Another approach to test analysis is called item response theory (IRT). With IRT, the relationship between an item and the trait measured by the total set of items is usually represented by an item characteristic curve, which is a nonlinear regression of the item on the measure of ability representing the total test. An index called the point-biserial correlation is sometimes computed in an IRT item analysis statistical program, but that correlation might not be exactly the same as the item-test correlation. The IRT index might be called an item-trait or an item-theta correlation, because the item is correlated with a measure of the ability estimated differently than based on the total score. Although this technical difference might exist, there is little substantial difference between the item-trait and the item-test correlations. The explanations in this entry

are intended to present information to help the user of either index.

*Darrell Sabers and Perman Gochyyev*

***See also*** Classical Test Theory; Coefficient Alpha; Internal Consistency Reliability; Item Analysis; Item Response Theory; Pearson Product-Moment Correlation Coefficient; Principal Components Analysis

**Further Readings**

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

# J

## JACKKNIFE

The jackknife or "leave one out" procedure is a cross-validation technique first developed by M. H. Quenouille to estimate the bias of an estimator. John Tukey then expanded the use of the jackknife to include variance estimation and tailored the name of *jackknife* because like a jackknife—a pocket knife akin to a Swiss army knife and typically used by Boy Scouts—this technique can be used as a "quick and dirty" replacement tool for a lot of more sophisticated and specific tools. Curiously, despite its remarkable influence on the statistical community, the seminal work of Tukey is available only from an abstract (which does not even mention the name of jackknife) and from an almost impossible to find unpublished note (although some of this note found its way into Tukey's complete work).

The jackknife estimation of a parameter is an iterative process. First the parameter is estimated from the whole sample. Then each element is, in turn, dropped from the sample and the parameter of interest is estimated from this smaller sample. This estimation is called a *partial estimate* (or also a *jackknife replication*). A *pseudovalue* is then computed as the difference between the whole sample estimate and the partial estimate. These pseudovalues reduce the (linear) bias of the partial estimate (because the bias is eliminated by the subtraction between the two estimates). The pseudovalues are then used in lieu of the original values to estimate the parameter of interest, and their standard deviation is used to estimate the parameter standard error, which can then be used for null hypothesis testing and for computing confidence intervals. The jackknife is strongly related to the bootstrap (i.e., the jackknife is often a linear approximation of the bootstrap), which is currently the main technique for computational estimation of population parameters.

As a potential source of confusion, a somewhat different (but related) method, also called jackknife, is used to evaluate the quality of the prediction of computational models built to *predict* the value of dependent variable(s) from a set of independent variable(s). Such models can originate, for example, from neural networks, machine learning, genetic algorithms, statistical learning models, or any other multivariate analysis technique. These models typically use a very large number of parameters (frequently *more* parameters than observations) and are therefore highly prone to overfitting (i.e., to be able to predict the data perfectly within the sample because of the large number of parameters but to be able to predict *new* observations poorly). In general, these models are too complex to be analyzed by current analytical techniques, and therefore, the effect of overfitting is difficult to evaluate directly. The jackknife can be used to

estimate the actual predictive power of such models by predicting the dependent variable values of each observation as if this observation were a new observation. To do so, the predicted value(s) of each observation is (are) obtained from the model built on the sample of observations minus the observation to be predicted. The jackknife, in this context, is a procedure that is used to obtain an unbiased prediction (i.e., a random effect) and to minimize the risk of overfitting.

## Definitions and Notations

The goal of the jackknife is to estimate a parameter of a population of interest from a random sample of data from this population. The parameter is denoted $\theta$, its estimate from a sample is denoted $T$, and its jackknife estimate is denoted $T^*$. The sample of $n$ observations (which can be univariate or multivariate) is a set denoted $\{X_1, \ldots, X_n, \ldots, X_n\}$. The sample estimate of the parameter is a function of the observations in the sample. Formally:

$$T=f(X_1,\ldots,X_n,\ldots,X_N). \tag{1}$$

An estimation of the population parameter obtained *without* the $n$th observation is called the $n$th *partial prediction* and is denoted $T_{-n}$. Formally:

$$T_{-n}=f(X_1,\ldots,X_{n-1},X_{n+1}\ldots,X_n). \tag{2}$$

A pseudovalue estimation of the $n$th observation is denoted $T_n^*$; it is computed as the difference between the parameter estimation obtained from the whole sample and the parameter estimation obtained without the $n$th observation. Formally:

$$T_n^* = NT - (N-1)T_{-n}. \tag{3}$$

The jackknife estimate of $\theta$, denoted $T^*$, is obtained as the mean of the pseudovalues. Formally:

$$T^* = \overline{T}_\bullet^* = \frac{1}{N}\sum_n^N T_n^*, \tag{4}$$

where $\overline{T}_\bullet^*$ is the mean of the pseudovalues. The variance of the pseudovalues is denoted $\hat{\sigma}^2_{T_n^*}$ and is obtained with the usual formula:

$$\hat{\sigma}^2_{T_n^*} = \frac{\sum\left(T_n^* - \overline{T}_\bullet^*\right)^2}{N-1}. \tag{5}$$

Tukey conjectured that the $\overline{T}_n^*$s could be considered as independent random variables. Therefore, the standard error of the parameter estimates, denoted $\hat{\sigma}^2_{T^*}$, could be obtained from the variance of the pseudovalues from the usual formula for the standard error of the mean as follows:

$$\hat{\sigma}_T^* = \sqrt{\frac{\hat{\sigma}^*_{T_n^*}}{N}} = \sqrt{\frac{\sum\left(T_n^* - \overline{T}_\bullet^*\right)^2}{N(N-1)}}. \tag{6}$$

This standard error can then be used to compute confidence intervals for the estimation of the parameter. Under the independence assumption, this estimation is distributed as a Student's $t$ distribution with $(N-1)$ degrees of freedom. Specifically a $(1-\alpha)$ confidence interval can be computed as

$$T^* \pm t_{\alpha,\nu}\hat{\sigma}_T^* \tag{7}$$

with $t_{\alpha,\nu}$ being the $\alpha$-level critical value of a Student's $t$ distribution with $\nu = N-1$ degrees of freedom.

### *Jackknife Without Pseudovalues*

Pseudovalues are important for understanding the inner working of the jackknife, but they are not computationally efficient. Alternative formulas using only the partial estimates can be used in lieu of the pseudovalues. Specifically, if $\overline{T}_\cdot$ denotes the mean of the partial estimates and $\hat{\sigma}_{T_{-n}}$ their standard deviation, then $T^*$ (cf. Equation 4) can be computed as

$$T^* = NT - (N-1)\overline{T}_\bullet \tag{8}$$

and $\hat{\sigma}_T^*$ (cf. Equation 6) can be computed as

$$\begin{aligned}\hat{\sigma}_{T^*} &= \sqrt{\frac{N-1}{N}\sum\left(T_{-n} - \overline{T}_\cdot\right)^2} \\ &= (N-1)\frac{\hat{\sigma}_{T_{-n}}}{\sqrt{N}}.\end{aligned} \tag{9}$$

### *Assumptions of the Jackknife*

Although the jackknife makes no assumptions about the shape of the underlying probability

distribution, it requires that the observations are independent of each other. Technically, the observations are assumed to be independent and identically distributed (i.e., in statistical jargon: i.i.d.). This means that the jackknife is not, in general, an appropriate tool for time-series data. When the independence assumption is violated, the jackknife underestimates the variance in the data set, which makes the data look more reliable than they actually are.

Because the jackknife eliminates the bias by subtraction (which is a linear operation), it works correctly only for statistics that are *linear* functions of the parameters or the data, and whose distribution is continuous or at least "smooth enough" to be considered as such. In some cases, linearity can be achieved by transforming the statistics (e.g., using a Fisher $Z$ transform for correlations or a logarithm transform for standard deviations), but some nonlinear or noncontinuous statistics, such as the median, will give very poor results with the jackknife no matter what transformation is used.

### *Bias Estimation*

The jackknife was originally developed by Quenouille as a nonparametric way to estimate and reduce the bias of an estimator of a population parameter. The bias of an estimator is defined as the difference between the expected value of this estimator and the true value of the population parameter. So formally, the bias, denoted ß, of an estimation $T$ of the parameter $\theta$ is defined as

$$ß = \mathbf{E}\{T\} - \theta, \tag{10}$$

with $\mathbf{E}\{T\}$ being the expected value of $T$.

The jackknife estimate of the bias is computed by replacing the expected value of the estimator [i.e., $\mathbf{E}\{T\}$] by the biased estimator (i.e., $T$) and by replacing the parameter (i.e., $\theta$) by the "unbiased" jackknife estimator (i.e., $T^*$). Specifically, the jackknife estimator of the bias, denoted $ß_{\text{jack}}$, is computed as

$$ß_{\text{jack}} = T - T^*. \tag{11}$$

### *Generalizing the Performance of Predictive Models*

Recall that the name *jackknife* refers to two related, but different, techniques (and this is sometimes a source of confusion). The first technique, presented in the preceding discussion, estimates population parameters and their standard error. The second technique evaluates the generalization performance of predictive models. In these models, predictor variables are used to predict the values of dependent variable(s). In this context, the problem is to estimate the quality of the prediction for *new* observations. Technically speaking, the goal is to estimate the performance of the predictive model as a *random effect* model. The problem of estimating the random effect performance for predictive models is becoming a crucial problem in domains such as, for example, bio-informatics and neuroimaging because the data sets used in these domains typically comprise a very large number of variables (often a much larger number of variables than observations—a configuration called the "small $N$, large $P$" problem). This large number of variables makes statistical models notoriously prone to overfitting.

In this context, the goal of the jackknife is to estimate how a model would perform when applied to *new* observations. This is done by dropping in turn each observation and fitting the model for the remaining set of observations. The model is then used to predict the left-out observation. With this procedure, each observation has been predicted as a new observation.

In some cases a jackknife can perform both functions, thereby generalizing the predictive model as well as finding the unbiased estimate of the parameters of the model.

## Example: Linear Regression

Suppose that we had performed a study examining the speech rate of children as a function of their age. The children's age (denoted $X$) would be used as a predictor of their speech rate (denoted $Y$). Dividing the number of words said by the time needed to say them would produce the speech rate (expressed in words per minute) of each child. The results of this (fictitious) experiment are shown in Table 1.

We will use these data to illustrate how the jackknife can be used to (a) estimate the regression parameters and their bias and (b) evaluate the generalization performance of the regression model. As a preliminary step, the data are analyzed by

**Table 1** Data From a Study Examining the Speech Rate of Children as a Function of Age

| | $X_n$ | $Y_n$ | $\hat{Y}_n$ | $\hat{Y}_n^*$ | $\hat{Y}_{jack,n}$ |
|---|---|---|---|---|---|
| 1 | 4 | 91 | 95.0000 | 9 4.9986 | 97.3158 |
| 2 | 5 | 96 | 96.2500 | 96.1223 | 96.3468 |
| 3 | 6 | 103 | 97.5000 | 97.2460 | 95.9787 |
| 4 | 9 | 99 | 101.2500 | 100.6172 | 101.7411 |
| 5 | 9 | 103 | 101.2500 | 100.6172 | 100.8680 |
| 6 | 15 | 108 | 108.7500 | 107.3596 | 111.3962 |

*Notes:* The independent variable is the age of the child ($X$). The dependent variable is the speech rate of the child in words per minutes ($Y$). The values of $\hat{Y}$ are obtained as $\hat{Y} = 90 + 1.25X$. $X_n$ is the value of the independent variable, $\hat{Y}_n$ is the value of the dependent variable, $Y_n$ is the predicted value of the dependent variable predicted from the regression, $\hat{Y}_n^*$ is the predicted value of the dependent variable predicted from the jackknife derived unbiased estimates, and $Y_{jack}$ is the predicted values of the dependent variable when each value is predicted from the corresponding jackknife partial estimates.

a standard regression analysis, and we found that the regression equation is equal to

$$\hat{Y} = a + bX = 90 + 1.25X. \quad (12)$$

The predicted values are given in Table 1. This regression model corresponds to a coefficient of correlation of $r$ = .8333 (i.e., the correlation between the $Y$s and the $\hat{Y}$s is equal to .8333).

### *Estimation of Regression Parameters and Bias*

In this section, the jackknife is used to estimate the intercept, the slope, and the value of the coefficient of correlation for the regression.

Each observation is dropped in turn and computed for the slope and the intercept, the partial estimates (denoted $b_{-n}$ and $a_{-n}$), and the pseudovalues (denoted $b_n^*$ and $a_n^*$). So, for example, when we drop the first observation, we use the observations 2 through 6 to compute the regression equation with the partial estimates of the slope and intercept as (cf. Equation 2):

$$\hat{Y}_{-1} = a_{-1} + b_{-1}X = 93.5789 + 0.9342X. \quad (13)$$

From these partial estimates, we compute a pseudovalue by adapting Equation 3 to the regression context. This gives the following jackknife pseudovalues for the $n$th observation:

$$\begin{aligned} a_n^* &= na - (n-1)a_{-n} \quad \text{and} \\ b_n^* &= nb - (n-1)b_{-n}, \end{aligned} \quad (14)$$

and for the first observation, this equation becomes

$$\begin{aligned} a_1^* &= 6 \times 1.25 - 5 \times 0.9342 = 2.8289 \quad \text{and} \\ b_1^* &= 6 \times 90 - 5 \times 93.5789 = 72.1053. \end{aligned} \quad (15)$$

Table 2 gives the partial estimates and pseudovalues for the intercept and slope of the regression. From this table, we can find that the jackknife estimates of the regression will give the following equation for the prediction of the dependent variable (the prediction using the jackknife estimates is denoted $\hat{Y}_n^*$):

$$\hat{Y}_n^* = a^* + b^* X = 90.5037 + 1.1237X. \quad (16)$$

The predicted values using the jackknife estimates are given in Table 1. It is worth noting that, for regression, the jackknife parameters are *linear* functions of the standard estimates. This implies that the values of $\hat{Y}_n^*$ can be perfectly predicted from the values of $\hat{Y}_n$. Specifically,

$$\hat{Y}_n^* = \left(a^* - a\frac{b^*}{b}\right) + \frac{b^*}{b}\hat{Y}_n. \quad (17)$$

Therefore, the correlation between the $\hat{Y}_n^*$ and the $Y_n$ is equal to one; this, in turn, implies that the correlation between the original data and the predicted values is the same for both $Y$ and $\hat{Y}_n^*$.

The estimation for the coefficient of correlation is slightly more complex because, as mentioned, the jackknife does not perform well with nonlinear statistics such as correlation. So, the values of $r$ are transformed using the Fisher $Z$ transform *prior* to jackknifing. The jackknife estimate is computed on these $Z$-transformed values, and the final value of the estimate of $r$ is obtained by using the inverse of the Fisher $Z$ transform (using $r$ rather than the transformed $Z$ values would lead to a gross overestimation of the correlation). Table 2 gives the partial estimates for the correlation, the $Z$-transformed values, and the $Z$-transformed pseudovalues. From Table 2, we find that the jackknife estimate of the $Z$-transformed coefficient of correlation is equal to $Z^* = 1.019$, which when

**Table 2** Partial Estimates and Pseudovalues for the Regression Example of Table 1

| | *Partial Estimates* | | | | *Pseudovalues* | | |
|---|---|---|---|---|---|---|---|
| *Observations* | $A_{-n}$ | $b_{-n}$ | $r_{-n}$ | $Z_{-n}$ | $a_n^*$ | $b_n^*$ | $Z^*$ |
| 1 | 93.5789 | 0.9342 | 8005 | 1.1001 | 72.1053 | 2.8289 | 1.6932 |
| 2 | 90.1618 | 1.2370 | 8115 | 1.1313 | 89.1908 | 1.3150 | 1.5370 |
| 3 | 87.4255 | 1.4255 | 9504 | 1.8354 | 102.8723 | 0.3723 | − 1.9835 |
| 4 | 90.1827 | 1.2843 | .8526 | 1.2655 | 89.0863 | 1.0787 | 0.8661 |
| 5 | 89.8579 | 1.2234 | .8349 | 1.2040 | 90.7107 | 1.3832 | 1.1739 |
| 6 | 88.1887 | 1.5472 | .7012 | 0.8697 | 99.0566 | − 0.2358 | 2.8450 |
| | $\overline{a}_\bullet$ | $\overline{b}_\bullet$ | — | $\overline{Z}_\bullet$ | $a^*$ | $b^*$ | $Z^*$ |
| | 89.8993 | 1.2753 | — | .2343 | 90.5037 | 1.1237 | 1.0219 |
| | | | | | *Jackknife Estimates* | | |
| *SD* | $\hat{\sigma}_{b_{-n}}$ | $\hat{\sigma}_{b_{-n}}$ | — | $\hat{\sigma}_{z_{-n}}$ | $\hat{\sigma}^*_{a_n}$ | $\hat{\sigma}^*_{b_n}$ | $\hat{\sigma}^*_{z_n}$ |
| | 2.1324 | 0.2084 | — | 0.3240 | 10.6622 | 1.0418 | 1.6198 |
| | | | | | *Jackknife Standard Deviations* | | |
| *SE* | — | — | — | — | $\hat{\sigma}_{a*} = \frac{\hat{\sigma}a^*_n}{\sqrt{n}}$ | $\hat{\sigma}_{b*} = \frac{\hat{\sigma}b^*_n}{\sqrt{n}}$ | $\hat{\sigma}_{z*} = \frac{\hat{\sigma}z^*_n}{\sqrt{n}}$ |
| | — | — | — | — | 4.3528 | 0.4253 | .6613 |
| | | | | | *Jackknife Standard Error* | | |

transformed back to a correlation, produces a value of the jackknife estimate for the correlation of $r^* = .7707$. Incidently, this value is very close to the value obtained with another classic alternative population unbiased estimate called the shrunken $r$, which is denoted $\tilde{r}$, and computed as

$$\tilde{r} = \sqrt{1 - \left[(1 - r^2)\frac{(N-1)}{(N-2)}\right]}$$
$$= \sqrt{1 - \left[(1 - .8333^2)\frac{5}{4}\right]} = .7862. \quad (18)$$

Confidence intervals are computed using Equation 7. For example, taking into account that the $\alpha = .05$ critical value for a Student's $t$ distribution for $\nu = 5$ degrees of freedom is equal to $t_{\alpha,\nu} = 2.57$, the confidence interval for the intercept is equal to

$$a^* \pm t_{\alpha,\nu}\hat{\sigma}_{a^*} = 90.5037 \pm 2.57 \times \frac{10.6622}{\sqrt{6}}$$
$$= 90.5037 \pm 2.57 \times 4.3528 \quad (19)$$
$$= 90.5037 \pm 11.1868.$$

The bias of the estimate is computed from Equation 11. For example, the bias of the estimation of the coefficient of correlation is equal to

$$\text{ß}_{\text{jack}}(r) = r - r^* = .8333 - .7707 = .0627. \quad (20)$$

The bias is positive, and this shows (as expected) that the coefficient of correlation overestimates the magnitude of the population correlation.

## Estimate of the Generalization Performance of the Regression

To estimate the generalization performance of the regression, we need to evaluate the performance of the model on *new* data. These data are supposed to be randomly selected from the same population as the data used to build the model. The jackknife strategy here is to predict each observation as a new observation; this implies that each observation is predicted from its *partial estimates* of the prediction parameter. Specifically, if we denote by $Y_{jack,n}$ the jackknife predicted value of the $n$th observation, the jackknife regression equation becomes

$$\hat{Y}_{jack,n} = a_{-n} + b_{-n}X_n. \quad (21)$$

So, for example, the first observation is predicted from the regression model built with observations 2 to 6; this gives the following predicting equation for $Y_{jack,1}$ (cf. Tables 1 and 2):

$$\hat{Y}_{jack,1} = a_{-1} + b_{-1}X_1 = 93.5789 + 0.9342 \times 4 = 97.3158. \quad (22)$$

The jackknife predicted values are listed in Table 1. The quality of the prediction of these jackknife values can be evaluated, once again, by computing a coefficient of correlation between the predicted values (i.e., the $Y_{jack,n}$) and the actual values (i.e., the $Y_n$). This correlation, denoted $r_{jack}$, for this example is equal to $r_{jack} = .6825$. It is worth noting that, in general, the coefficient $r_{jack}$ is *not* equal to the jackknife estimate of the correlation $r^*$ (which, recall, is in our example equal to $r^* = .7707$).

*Hervé Abdi and Lynne J. Williams*

***See also*** Bias; Bivariate Regression; Bootstrapping; Coefficients of Correlation, Alienation, and Determination; Pearson Product-Moment Correlation Coefficient; $R^2$; Reliability; Standard Error of Estimate

**Further Readings**


Abdi, H., Edelman, B., Valentin, D., & Dowling, W. J. (2009). *Experimental design and analysis for psychology.* Oxford, UK: Oxford University Press.

Bissell, A. F., & Ferguson, R. A. (1975). The jackknife—toy, tool or two-edged weapon? *The Statistician,* 24, 79–100.

Diaconis, P., & Efron, B. (1983). Computer-intensive methods in statistics. *Scientific American,* 116–130.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician* 37, 36–48.

Gentle, J. E. (1998). *Elements of computational statistics*. New York, NY: Springer.

Kriegeskorte, K., Simmons, W. K., Bellgowan, P. S. F., & Baker, C. I. (2009). Circular analysis in systems neuroscience: The dangers of double dipping. *Nature Neuroscience,* 12, 535–540.

Krzanowski, W. J., & Radley, D. (1989). Nonparametric confidence and tolerance regions in canonical variate analysis. *Biometrics,* 45, 1163–1173.

Hinkley, D. V. (1983). Jackknife methods. In Johnshon, N. L., Kotz, S., & Read, C. B. (Eds.). *Encyclopedia of statistical sciences* (Vol. 4, pp. 280–287). New York: Wiley.

Manly, B. F. J. (1997). *Randomization, bootstrap, and Monte Carlo methods in biology* (2nd ed.). New York: Chapman & Hall.

Miller, R. G. (1974). The jackknife: A review. *Biometrika,* 61, 1–17.

Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika,* 43, 353–360.

Shao, J., & Tu, D. (1995). *The jackknife and the bootstrap*. New York: Springer-Verlag.

Tukey, J. W. (1958). Bias and confidence in not quite large samples (abstract). *Annals of Mathematical Statistics,* 29, 614.

Tukey, J. W. (1986). The future of processes of data analysis. In *The collected works of John W. Tukey* (Vol. IV, pp. 517–549). New York: Wadsworth.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in *f*MRI studies of emotion, personality, and social cognition. *Perspectives in Psychological Sciences,* 4, 274–290.


# John Henry Effect

The term *John Henry effect* was coined to explain the unexpected outcome of an experiment caused by the control group's knowledge of its role within the experiment. The control group's perceived role as a baseline or a comparison group to the experimental condition, specifically one testing an innovative technology, can cause the control group to behave in an unnatural way to outperform the new technology. The group's knowledge of its position within the experiment as a baseline comparison causes the group to perform differently and, often more specifically, better than usual, eliminating the effect of the experimental manipulation. Deriving its name from the folktale the "Ballad of John Henry," the John Henry effect is similar to the Hawthorne effect in which participant behavior changes as a result of the participants' knowledge that they are being observed or studied. This change in participant behavior can confound the experiment rendering the results inaccurate or misleading.

The effect was studied and explained by education researcher Robert Heinich after his review of several studies that compared the effects of television instruction with those of standard classroom

teaching. Heinich noted that many of these studies demonstrated insignificant differences between control and experimental groups and often included results in which the control group outperformed the experimental condition. He was one of the first researchers to acknowledge that the validity of these experiments was compromised by the control groups' knowledge of their role as a control or baseline comparison group. Comparing the control groups with the title character from the "Ballad of John Henry," Heinich described how a control group might exert extra effort to compete with or even outperform its comparison group.

In the "Ballad," title character John Henry works as a rail driver whose occupation involves hammering spikes and drill bits into railroad ties to lay new tracks. John Henry's occupation is threatened by the invention of the steam drill, a machine designed to do the same job in less time. The "Ballad of John Henry" describes an evening competition in which Henry competes with the steam drill one on one and defeats it by laying more track. Henry's effort to outperform the steam drill causes a misleading result, however, because although he did in fact win the competition, his overexertion causes his death the next day.

Heinich's use of the folktale compares the performance by a control group with the performance of Henry in which an unexpected result develops from the group's overexertion or out-of-the-ordinary performance. With both Henry and the control group, the fear of being replaced incites the spirit of competition and leads to an inaccurate depiction of the differences in performance by the control and experimental groups.

The effect was later studied extensively by Gary Saretsky, who expanded on the term's definition by pointing out the roles that both competition and fear play in producing the effect. In most cases, the John Henry effect is perceived as a control group's resultant behavior to the fear of being outperformed or replaced by the new strategy or novel technology.

*Samantha John*

***See also*** Control Group; Hawthorne Effect

**Further Readings**

Heinich, R. (1970). *Technology & the management of instruction.* Washington, DC: Association for Educational Communications and Technology.

Saretsky, G. (1975). The John Henry Effect: Potential confounder of experimental vs. control group approaches to the evaluation of educational innovations. Presented at the annual meeting of the American Educational Research Association, Washington, DC.

# K

## Kolmogorov–Smirnov Test

The Kolmogorov–Smirnov (KS) test is one of many goodness-of-fit tests that assess whether univariate data have a hypothesized continuous probability distribution. The most common use is to test whether data are normally distributed. Many statistical procedures assume that data are normally distributed. Therefore, the KS test can help validate use of those procedures. For example, in a linear regression analysis, the KS test can be used to test the assumption that the errors are normally distributed. However, the KS test is not as powerful for assessing normality as other tests such as the Shapiro–Wilk, Anderson–Darling, and Bera–Jarque tests that are specifically designed to test for normal distributions. That is, if the data are not normal, the KS test will erroneously conclude that they are normal more frequently than will the other three mentioned tests. Yet the KS test is better in this regard than the widely used chi-square goodness-of-fit test. Nevertheless, the KS test is valid for testing data against *any* specified continuous distribution, not just the normal distribution. The other three mentioned tests are not applicable for testing non-normal distributions. Moreover, the KS test is distribution free, which means that the same table of critical values might be used—whatever the hypothesized continuous distribution, normal or otherwise.

This entry discusses the KS test in relation to estimating parameters, multiple samples, and goodness-of-fit tests. An example illustrating the application and evaluation of a KS test is also provided.

### Estimating Parameters

Most properties of the KS test have been developed for testing completely specified distributions. For example, one tests not just that the data are normal, but more specifically that the data are normal with a certain mean and a certain variance. If the parameters of the distribution are not known, it is common to estimate parameters in order to obtain a completely specified distribution. For example, to test whether the errors in a regression have a normal distribution, one could estimate the error variance by the mean-squared error and test whether the errors are normal with a mean of zero and a variance equal to the calculated mean-squared error. However, if parameters are estimated in the KS test, the critical values in standard KS tables are incorrect and substantial power can be lost. To permit parameter estimation in the KS test, statisticians have developed corrected tables of critical values for testing special distributions. For example, the adaptation of the KS test for testing the normal distribution with estimated mean and variance is called the Lilliefors test.

### Multiple-Sample Extensions

The KS test has been extended in other ways. For example, there is a two-sample version of the KS

test that is used to test whether two separate sets of data have the same distribution. As an example, one could have a set of scores for males and a set of scores for females. The two-sample KS test could be used to determine whether the distribution of male scores is the same as the distribution of female scores. The two-sample KS test does not require that the form of the hypothesized common distribution be specified. One does not need to specify whether the distribution is normal, exponential, and so on, and no parameters are estimated. The two-sample KS test is distribution free, so just one table of critical values suffices. The KS test has been extended further to test the equality of distributions when the number of samples exceeds two. For example, one could have scores from several different cities.

## Goodness-of-Fit Tests

In all goodness-of-fit tests, there is a null hypothesis that states that the data have some distribution (e.g., normal). The alternative hypothesis states that the data do not have that distribution (e.g., not normal). In most empirical research, one hopes to conclude that the data have the hypothesized distribution. But in empirical research, one customarily sets up the research hypothesis as the alternative hypothesis. In goodness-of-fit tests, this custom is reversed. The result of the reversal is that a goodness-of-fit test can provide only a weak endorsement of the hypothesized distribution. The best one can hope for is a conclusion that the hypothesized distribution *cannot be rejected,* or that the data are *consistent with the hypothesized distribution.* Why not follow custom and set up the hypothesized distribution as the alternative hypothesis? Then rejection of the null hypothesis would be a strong endorsement of the hypothesized distribution at a specified low Type I error probability. The answer is that it is too hard to disprove a negative. For example, if the hypothesized distribution were standard normal, then the test would have to disprove *all* other distributions. The Type I error probability would be 100%.

Like all goodness-of-fit tests, the KS test is based on a measure of disparity between the empirical data and the hypothesized distribution. If the disparity exceeds a critical cutoff value, the hypothesized distribution is rejected. Each goodness-of-fit test uses a different measure of disparity. The KS test uses the maximum distance between the empirical distribution function of the data and the hypothesized distribution. The following example clarifies ideas.

## Example

Suppose that one wishes to test whether the 50 randomly sampled data in Table 1 have a standard normal distribution (i.e., normal with mean $= 0$ and variance $= 1$). The smooth curve in Figure 1 shows the cumulative distribution function (CDF) of the standard normal distribution. That is, for each $x$ on the horizontal axis, the smooth curve shows the standard normal probability less than or equal to $x$. These are the values displayed in every table of standard normal probabilities found in statistics textbooks. For example, if $x = 2$, the smooth curve has a value of 0.9772, indicating that 97.72% of the values in the distribution are found below 2 standard deviations above the mean, and only 2.28% of the values are more than 2 standard deviations above the mean. The jagged line in the figure shows the empirical distribution function (EDF; i.e., the proportion of the 50 data less than or equal to each $x$ on the horizontal axis). The EDF is 0 below the minimum data value and is 1 above the largest data value and steps up by $1/50 = 0.02$ at each data value from left to right.

If the 50 data come from a standard normal distribution, then the smooth curve and the jagged line should be close together because the empirical proportion of data less than or equal to each $x$ should be close to the proportion of the true distribution less than or equal to each $x$. If the true distribution is standard normal, then any disparity or gap between the two lines should be attributable to the random variation of sampling, or to the discreteness of the 0.02 jumps at each data value. However, if the true data distribution is not standard normal, then the true CDF and the standard normal smooth curve of Figure 1 will differ. The EDF will be closer to the true CDF than to the smooth curve. And a persistent gap will open between the two curves of Figure 1, provided the sample size is sufficiently large. Thus, if there is only a small gap between the two curves of the figure, then it is plausible that the data come from

**Table 1** Example of KS Test (50 Randomly Sampled Data, Arranged in Increasing Order)

| | | | | |
|---|---|---|---|---|
| −1.7182 | −0.9339 | −0.2804 | 0.2694 | 1.0030 |
| −1.7144 | −0.9326 | −0.2543 | 0.3230 | 1.0478 |
| −1.6501 | −0.8866 | −0.2093 | 0.4695 | 1.0930 |
| −1.5493 | −0.6599 | −0.1931 | 0.5368 | 1.2615 |
| −1.4843 | −0.5801 | −0.1757 | 0.5686 | 1.2953 |
| −1.3246 | −0.5559 | −0.1464 | 0.7948 | 1.4225 |
| −1.3173 | −0.4403 | −0.0647 | 0.7959 | 1.6038 |
| −1.2435 | −0.4367 | −0.0594 | 0.8801 | 1.6379 |
| −1.1507 | −0.3205 | 0.0786 | 0.8829 | 1.6757 |
| −0.9391 | −0.3179 | 0.1874 | 0.9588 | 1.6792 |

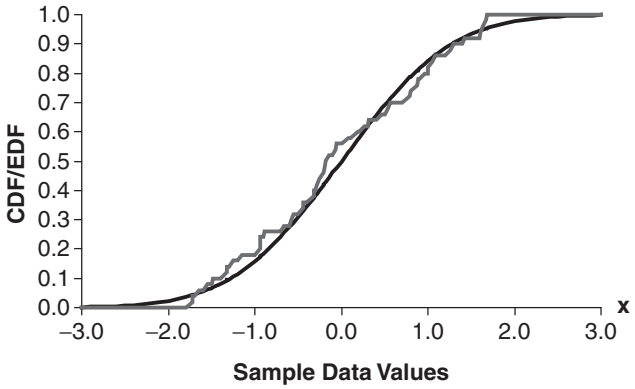



**Figure 1** Example of KS Test

*Notes:* The smooth curve is the CDF of hypothesized standard normal distribution. The jagged line is the EDF of 50 randomly sampled data in Table 1. The KS statistic is the maximum vertical distance between the curve and the line.

a standard normal distribution. However, if the gap is sufficiently large, then it is implausible that the data distribution is standard normal.

### *Intuition for the KS test*

This intuition is the basis for the KS test: Reject the hypothesis that the data are standard normal if the gap in Figure 1 is greater than a critical value; declare the data consistent with the hypothesis that the data are standard normal if the gap is less than the critical value. To complete the test, it is necessary to determine how to compute the gap and the critical value. The KS test measures the gap by the maximum vertical difference between the two lines. The maximum distance is not the only way to measure the gap. The area between the lines and square of the area betweenthe lines are among the measures that are used in other goodness-of-fit tests. In most cases, statistical software running on a computer will probably calculate both the value of the KS test and the critical value or $p$ value. If the computation of the KS test is done by hand, it is necessary to exercise some care because of the discontinuities or jumps in the EDF at each data value.

### *Calculation of the KS Statistic*

The KS statistic that measures the gap can be calculated in general by the formula

$$D_n = \max(D_n^+, D_n^-),$$

in which

$$D_n^+ = \max_{1 \le i \le n} [i/n - F_0(x_i)]$$

and

$$D_n^- = \max_{1 \le i \le n} [F_0(x_i) - (i-1)/n].$$

In these formulas, $D_n$ is the value of the KS statistic based on $n$ sample data that have been ordered ($x_1 \le x_2 \le \cdots \le x_n$); $F_0(x_i)$ is the value of the hypothesized CDF at $x_i$ (the smooth curve in the figure); and $i/n$ is the EDF (the proportion of data less than or equal to $x_i$). The statistics $D_n^+$ and $D_n^-$ are the basis of one-sided KS tests, in which the alternative hypothesis is that every percentile of $F_0$ exceeds (or lies below) every corresponding percentile of the true distribution $F$. Such one-sided KS tests have fairly good power.

### *Critical Values for the KS Test*

Formally, the null and alternative hypotheses for the KS test can be stated as

$$H_0 : F(x) = F_0(x) \text{ for all x}$$

versus

$$H_1 : F(x) \neq F_0(x) \text{ for at least one } x.$$

The hypotheses say that either the true CDF $[F(x)]$ is identical to the hypothesized CDF $[F_0(x)]$

at all $x$, or it is not. Statisticians have produced tables of critical values for small sample sizes. If the sample size is at least moderately large (say, $n \geq 35$), then asymptotic approximations might be used. For example, $P(D_n > 1.358/\sqrt{n}) = 0.05$ and $P(D_n > 1.224/\sqrt{n}) = 0.05$ for large $n$. Thus, $1.358/\sqrt{n}$ is an approximate critical value for the KS test at the 0.05 level of significance, and $1.224/\sqrt{n}$ is an approximate critical value for the KS test at the 0.10 level of significance.

In the example, the maximum gap between the EDF and the hypothesized standard normal distribution of Figure 1 is 0.0866. That is, the value of the KS statistic $D_n$ is 0.0866. Because $n = 50$, the critical value for a test at the 0.10 significance level is approximately $1.224/\sqrt{50} = 0.1731$. Because $D_n = 0.0866 < 0.1731$, then it could be concluded that the data are consistent with the hypothesis of a standard normal distribution at the 0.10 significance level. In fact, the $p$ value is about 0.8472.

However, the data in Table 1 are in fact drawn from a true distribution that is not standard normal. The KS test incorrectly accepts (fails to reject) the null hypothesis that the true distribution is standard normal. The KS test commits a Type II error. The true distribution is uniform on the range $(-1.7321, +1.7321)$. This uniform distribution has a mean of 0 and a variance of 1, just like the standard normal. The sample size of 50 is insufficient to distinguish between the hypothesized standard normal distribution and the true uniform distribution with the same mean and variance. The maximum KS gap between the true uniform distribution and the hypothesized standard normal distribution is only 0.0572. Because the critical value for the KS test using 50 data and a 0.10 significance level is about three times the true gap, it is very unlikely that an empirical gap of the magnitude required to reject the hypothesized standard normal distribution can be obtained. A much larger data set is required. The critical value $(1.224/\sqrt{n})$ for a test at the 0.10 significance level must be substantially smaller than the true gap (0.0572) for the KS test to have much power.

### *Evaluation*

Several general lessons can be drawn from this example. First, it is difficult for the KS test to distinguish small differences between the hypothesized distribution and the true distribution unless a substantially larger sample size is used than is common in much empirical research. Second, failure to reject the hypothesized distribution might be only a weak endorsement of the hypothesized distribution. For example, if one tests the hypothesis that a set of regression residuals has a normal distribution, one should perhaps not take too much comfort in the failure of the KS test to reject the normal hypothesis. Third, if the research objective can be satisfied by testing a small set of specific parameters, it might be overkill to test the entire distribution. For example, one might want to know whether the mean of male scores differs from the mean of female scores. Then it would probably be better to test equality of means (e.g., by a $t$ test) than to test the equality of distributions (by the KS test). The hypothesis of identical distributions is much stronger than the hypothesis of equal means and/or variances. As in the example, two distributions can have equal means and variances but not be identical. To have identical distributions means that *all possible* corresponding pairs of parameters are equal. The KS test has some sensitivity to *all* differences between distributions, but it achieves that breadth of sensitivity by sacrificing sensitivity to differences in specific parameters. Fourth, if it is not really important to distinguish distributions that differ by small gaps (such as 0.0572)—if only large gaps really matter—then the KS test might be quite satisfactory. In the example, this line of thought would imply researcher indifference to the shapes of the distributions (uniform vs. normal)—the uniform distribution on the range $(-1.7321, +1.7321)$ would be considered "close enough" to normal for the intended purpose.

*Thomas W. Sager*

***See also*** Distribution; Nonparametric Statistics

**Further Readings**

Conover, W. J. (1999). *Practical nonparametric statistics* (3rd ed.). New York: Wiley.

Hollander, M., & Wolfe, D. A. (1999). *Nonparametric statistical methods* (2nd ed.). New York: Wiley-Interscience.

Khamis, H. J. (2000). The two-stage $\delta$-corrected Kolmogorov-Smirnov test. *Journal of Applied Statistics,* 27, 439–450.

Lilliefors, H. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association, 62,* 399–402.

Stephens, M. A. (1974). EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association, 69,* 730–737.

Stephens, M. A. (1986). Tests based on EDF statistics. In R. B. D'Agostino, & M. A. Stevens (Eds.), *Goodness-of-fit techniques.* New York: Marcel Dekker.

# KR-20

KR-20 (Kuder–Richardson Formula 20) is an index of the internal consistency reliability of a measurement instrument, such as a test, questionnaire, or inventory. Although it can be applied to any test item responses that are dichotomously scored, it is most often used in classical psychometric analysis of psychoeducational tests and, as such, is discussed with this perspective.

Values of KR-20 generally range from 0.0 to 1.0, with higher values representing a more internally consistent instrument. In very rare cases, typically with very small samples, values less than 0.0 can occur, which indicates an extremely unreliable measurement. A rule-of-thumb commonly applied in practice is that 0.7 is an acceptable value or 0.8 for longer tests of 50 items or more. Squaring KR-20 provides an estimate of the proportion of score variance not resulting from error. Measurements with KR-20 $< 0.7$ have the majority of score variance resulting from error, which is unacceptable in most situations.

Internal consistency reliability is defined as the consistency, repeatability, or homogeneity of measurement given a set of item responses. Several approaches to reliability exist, and the approach relevant to a specific application depends on the sources of error that are of interest, with internal consistency being appropriate for error resulting from differing items.

KR-20 is calculated as

$$KR20 = \frac{K}{K-1}\left[1 - \frac{\sum_{i=1}^{K} p_i q_i}{\sigma_X^2}\right], \qquad (1)$$

where $K$ is the number of $i$ items or observations, $p_i$ is the proportion of responses in the keyed direction for item $i, q_i = 1 - p_i$, and $\sigma_X^2$ is the variance of the raw summed scores. Therefore, KR-20 is a function of the number of items, item difficulty, and the variance of examinee raw scores. It is also a function of the item-total correlations (classical discrimination statistics) and increases as the average item-total correlation increases.

KR-20 produces results equivalent to coefficient $\alpha$, which is another index of internal consistency, and can be considered a special case of $\alpha$. KR-20 can be calculated only on dichotomous data, where each item in the measurement instrument is cored into only two categories. Examples of this include true/false, correct/incorrect, yes/no, and present/absent. Coefficient $\alpha$ also can be calculated on polytomous data, that is, data with more than two levels. A common example of polytomous data is a Likert-type rating scale.

Like $\alpha$, KR-20 can be described as the mean of all possible split-half reliability coefficients based on the Flanagan–Rulon approach of split-half reliability. An additional interpretation is derived from Formula 1: The term $p_i q_i$ represents the variance of each item. If this is considered error variance, then the sum of the item variances divided by the total variance in scores presents the proportion of variance resulting from error. Subtracting this quantity from 1 translates it into the proportion of variance *not* resulting from error, assuming there is no source of error other than the random error present in the process of an examinee responding to each item.

G. Frederic Kuder and Marion Richardson also developed a simplification of KR-20 called KR-21, which assumes that the item difficulties are equivalent. KR-21 allows us to substitute the mean of the $p_i$ and $q_i$ into Formula 1 for $p_i$ and $q_i$, which simplifies the calculation of the reliability.

An important application of KR-20 is the calculation of the classical standard error of measurement (*SEM*) for a measurement. The *SEM* is

$$SEM = s_X\sqrt{1 - KR20}, \qquad (2)$$

where $s_X$ is the standard deviation of the raw scores in the sample. Note that this relationship is inverse; with $s_X$ held constant, as KR-20 increases,

the SEM decreases. Within classical test theory, having a more reliable measurement implies a smaller SEM for all examinees.

The following data set is an example of the calculation of KR-20 with 5 examinees and 10 items.

| | *Item* | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *Person* | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* | *9* | *10* | X |
| 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 9 |
| 2 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 7 |
| 3 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 6 |
| 4 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 5 |
| 5 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 8 |
| $p_i$: | 0.8 | 0.2 | 1.0 | 0.8 | 0.4 | 0.8 | 0.8 | 0.4 | 0.8 | 0.8 | |
| $q_i$: | 0.2 | 0.8 | 0.0 | 0.2 | 0.6 | 0.2 | 0.2 | 0.6 | 0.2 | 0.2 | |
| $p_i \times q_i$ | 0.16 | 0.16 | 0.0 | 0.16 | 0.24 | 0.16 | 0.16 | 0.24 | 0.16 | 0.16 | |

The variance of the raw scores ($X$) in the final column is 2.5, resulting in

$$\text{KR-20} = \frac{10}{10-1}\left[1 - \frac{1.6}{2.5}\right] = 0.4$$

as the KR-20 estimate of reliability.

*Nathan A. Thompson*

***See also*** Classical Test Theory; Coefficient Alpha; Internal Consistency Reliability; Reliability

**Further Readings**

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78,* 98–104.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.

de Gruijter, D. N. M., & van der Kamp, L. J. T. (2007). *Statistical test theory for the behavioral sciences*. Boca Raton, FL: CRC Press.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2,* 151–160.

# Krippendorff's Alpha

Krippendorff's $\alpha$ (alpha) is a general statistical measure of agreement among observers, measuring devices, or coders of data, designed to indicate their reliability. As a general measure, it is applicable to data on various levels of measurement (metrics) and includes some known coefficients as special cases. As a statistical measure, it maps samples from a population of data into a single chance corrected coefficient, a scale, indicating the extent to which the population of data can be relied on or trusted in subsequent analyses. Alpha equates reliability with the reproducibility of the data-generating process, measured by the agreement on what the data in question refer to or mean. Typical applications of $\alpha$ are content analyses where volumes of text need to be read and categorized, interview responses that require scaling or ranking before they can be treated statistically, or estimates of political or economic variables.

## Reliability Data

Data are considered reliable when researchers have reasons to be confident that their data represent real phenomena in the world outside their project, or are not polluted by circumstances that are extraneous to the process designed to generate them. This confidence erodes with the emergence of disagreements, for example, among human coders regarding how they judge, categorize, or score given units of analysis, in the extreme, when their accounts of what they see or read is random. To establish reliability requires duplications of the data-making efforts by an ideally large number of coders. Figure 1 represents reliability data in their most basic or canonical form, as a matrix of $m$ coders by $r$ units, containing the values $c_{iu}$ assigned

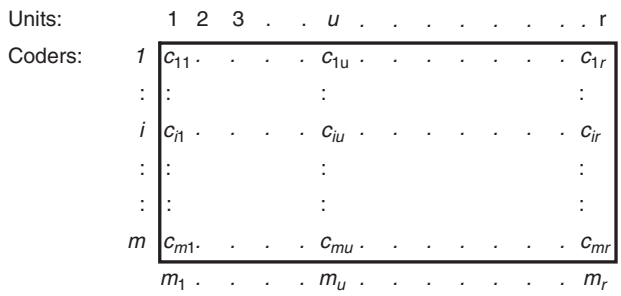



**Figure 1** Canonical Form of Reliability Data

by coder $i$ to unit $u$. The total number of pairable values $c$ is

$$n = \sum_u m_u | m_u \geq 2,$$

where $m_u$ is the number of coders evaluating unit $u$.

For data to serve reliability assessments, it is necessary, moreover, that (a) units are freely permutable and (b) representative of the data whose reliability is in question; and that (c) coders work independent of each other and (d) must be sufficiently common to be found where the data-making process might be replicated or data are to be added to an existing project.

## Alpha

The general form of Krippendorff's $\alpha$ is

$$\alpha = 1 - \frac{D_o}{D_e},$$

where $D_o$ is the observed disagreement and $D_e$ is the disagreement expected when the correlation between the units coded and the values used by coders to describe them is demonstrably absent. This conception of chance is uniquely tied to data-making processes.

When agreement is without exception, $D_o = 0$, $\alpha = 1$, and data reliability is considered perfect. When $D_o = D_e$, $\alpha = 0$, and reliability is considered absent. In statistical reality, $\alpha$ might be negative, leading to these limits:

$$1 \geq \alpha \geq 0 \begin{cases} \pm \text{Sampling error} \\ - \text{Systematic disagreement} \end{cases}.$$

Small sample sizes might cause the zero value of $\alpha$ to be a mere approximation. The occurrence of systematic disagreement can drive $\alpha$ below zero. The latter should not occur when coders follow the same coding instruction and work independently of each other as is required for generating proper reliability data.

In terms of the reliability data in Figure 1, $\alpha$ is defined—for conceptual clarity expressed here without algebraic simplifications—by

$$\alpha_{\text{metric}} = 1 - \frac{D_o}{D_e} = 1 - \frac{\sum_{u=1}^{u=r} \frac{m_u}{n} \sum_{i-1}^{i=m} \sum_{j=1}^{j=m} \frac{{}_{\text{metric}}\delta^2_{c_{iu}k_{ju}}}{m_u(m_u-1)}}{\sum_{c=1}^{c=n} \sum_{k=1}^{k=n} \frac{{}_{\text{metric}}\delta^2_{ck}}{n(n-1)}}.$$

In this expression, the denominator $D_e$ is the average difference ${}_{\text{metric}}\delta^2_{ck}$ between the $n(n-1)$ pairs of values $c$ and $k$ in the $n$ entries in Table 1 regardless of their occurrence in units or which coder contributed them, excluding values in units with $m_u \leq 1$. The number $n(n-1)$ excludes the pairing of values with themselves. The numerator $D_o$ first averages the differences ${}_{\text{metric}}\delta^2_{c_{iu}k_{iu}}$ between all $m_u(m_u-1)$ pairable values $c_{iu}$ and $k_{ju}$ that can be formed within the $m_u \geq 2$ values in units $u$ and then averages this difference over all units $u$.

All differences have the properties ${}_{\text{metric}}\delta^2_{ck} \geq 0$, ${}_{\text{metric}}\delta^2_{ck} = {}_{\text{metric}}\delta^2_{kc}$, ${}_{\text{metric}}\delta^2_{cc} = {}_{\text{metric}}\delta^2_{kk} = 0$, and respond to the metric or level of measurement of the data involved. The difference functions provided by $\alpha$ are

${}_{\text{nominal}}\delta^2_{ck} = \begin{cases} 0 \text{ iff c} = k \\ 1 \text{ iff c} \neq k \end{cases}$, where $c$ and $k$ are names;

${}_{\text{ordinal}}\delta^2_{ck} = \left( \sum_{g=c}^{g=k} n_g - \frac{n_c + n_k}{2} \right)^2$, where $n_g$ is the number of ranks $g$, used by all coders;

${}_{\text{interval}}\delta^2_{ck} = (c-k)^2$, where $c$ and $k$ are the values of an interval scale;

${}_{\text{ratio}}\delta^2_{ck} = \left( \frac{c-k}{c+k} \right)^2$, where $c$ and $k$ are absolute values;

${}_{\text{polar}}\delta^2_{ck} = \frac{(c-k)^2}{(c+k-2c_{\min})(2c_{\max}-c-k)}$, where $c_{\min}$ and $c_{\max}$ are the extreme bipolar values of a scale;

${}_{circular}\delta^2_{ck} = \left( \sin\left[ 180 \frac{c-k}{U} \right] \right)^2$, where $U$ is the range of values in a circle and the arguments of *sin* are degrees.

Thus, $\alpha_{\text{nominal}}$ measures agreements in nominal data or categories, $\alpha_{\text{ordinal}}$ in ordinal data or rank orderings, $\alpha_{\text{interval}}$ in interval data or scale points,

$\alpha_{\text{ratio}}$ in ratio data such as proportions or absolute numbers, $\alpha_{\text{polar}}$ in data recorded in bipolar opposite scales, and $\alpha_{\text{circular}}$ in data whose values constitute a closed circle or recursions.

The earlier expression for $\alpha$ in terms of Figure 1 is computationally inefficient and can be simplified in terms of a conceptually more convenient coincidence matrix representation, shown in Figure 2, summing the reliability data in Figure 1 as indicated. Coincidence matrices take advantage of the fact that reliability does not depend on the identity of coders, only on whether pairs of values match and what it means if they do not match, and on estimates of how these values are distributed in the population of data whose reliability is in question. Therefore, they tabulate coincidences without reference to coders. Coincidence matrices should not be confused with the more familiar contingency matrices that cross-tabulate units of analysis as judged or responded to by two coders (not the values they jointly generate).

Categories: 1 . k . .

| | 1 | . | k | . | . | |
|---|---|---|---|---|---|---|
| 1 | $n_{11}$ | . | $n_{1k}$ | . | . | $n_{1\cdot}$ |
| . | . | . | . | . | . | . |
| c | $n_{c1}$ | . | $n_{ck}$ | . | . | $n_{c\cdot}$ |
| . | . | . | . | . | . | . |
| . | . | . | . | . | . | . |
| | $n_{\cdot 1}$ | . | $n_{\cdot k}$ | . | . | $n_{\cdot\cdot}$ = Number of values used by all coders |

Where: $n_{ck} = \sum_u \frac{\text{number of } c-k \text{ pairs of values in unit } u}{m_u - 1}$

**Figure 2** Coincidence Matrix Representation of Reliability Data

By contrast, the cell contents $n_{ck}$ of coincidence matrices are the frequencies of $c_u - k_u$ pairs of values found in units $u$, weighted by $(m_u - 1)$ to ensure that each pairable value contributes exactly one to the matrix. Coincidence matrices contain perfectly matching values, $n_{cc}$, in their diagonal and are symmetrical around that diagonal, $n_{ck} = n_{kc}$. Their marginal sums $n_{c\cdot} = n_{\cdot c}$ enumerate the values $c$ used by all coders. Their totals, $n_{\cdot\cdot} = n \leq mr$, are equal to $mr$ when the table in Figure 1 is fully occupied, and less then $mr$ when values are missing, including values that are not pairable within units. The distribution of the marginal sums, $n_{c\cdot}$ and $n_{\cdot c}$, is the best estimate of the otherwise unknown distribution of values in the population of data whose reliability is in question.

In coincidence matrix terms, $\alpha$ becomes

$$\alpha_{\text{metric}} = 1 - \frac{\sum_c \sum_k n_{ck}\,\text{metric}\,\delta^2_{ck}}{\sum_c \sum_k \frac{n_{c\cdot} n_{\cdot k}}{n_{\cdot\cdot} - 1}\,\text{metric}\,\delta^2_{ck}}.$$

This expression might be simplified with reference to particular metrics, for example, for nominal and binary data:

$$\alpha_{\text{nominal}} = 1 - \frac{\sum_c \sum_{k \neq c} n_{ck}}{\sum_c \sum_{k \neq c} \frac{n_{c\cdot} n_{\cdot k}}{n_{\cdot\cdot} - 1}} = \frac{\sum_c n_{cc} - \sum_c \frac{n_{c\cdot}(n_{c\cdot} - 1)}{n_{\cdot\cdot} - 1}}{n_{\cdot\cdot} - \sum_c \frac{n_{c\cdot}(n_{c\cdot} - 1)}{n_{\cdot\cdot} - 1}},$$

$$\alpha_{\text{binary}} = 1 - (n_{\cdot\cdot} - 1)\frac{n_{c \neq k}}{n_{c\cdot} n_{\cdot k}}.$$

It should be mentioned that the family of $\alpha$ coefficients also includes versions for coding units with multiple values as well as for unitizing continua, for example, of texts taken as character strings or tape recordings. These are not discussed here.

## An Example

Consider the example in Figure 3 of three coders, each assigning one of four values to most of the 11 units.

On the left of Figure 3 the values of the reliability data are tabulated in their canonical form. Coders $h$ and $j$ code only 9 out of the 11 units that are attended to by coder $i$. Excluding unit 11, which does not contain pairable values, $m_{11} \leq 1$, all $n = n_{\cdot\cdot} = 28$ pairable values are found tabulated in the coincidence matrix on the right of Figure 3. For nominal data, $\alpha_{\text{nominal}} = 0.624$. If $\alpha$ is interpreted as the proportion of values that perfectly distinguish among the given units, the remainder resulting from chance, this might be seen in the canonical form of the reliability data. In the first six units, one finds 14 out of 28 values in perfect agreement. Very few of them could be the results of

| Units $u$: | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Coder $h$: | 2 | 3 | | 4 | 4 | 2 | 2 | 4 | 4 | 2 | |
| $i$: | 2 | 3 | 1 | 4 | 4 | 2 | 1 | 3 | 3 | 3 | 3 |
| $j$: | 2 | 3 | 1 | 4 | 4 | 2 | 1 | 4 | 3 | | |
| $m_u$: | 3 | 3 | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 2 | 1 |

| Categories $c$: | 1 | 2 | 3 | 4 | |
|---|---|---|---|---|---|
| $k$: 1 | 3 | 1 | | | 4 |
| 2 | 1 | 6 | 1 | | 8 |
| 3 | | 1 | 4 | 2 | 7 |
| 4 | | | 2 | 7 | 9 |
| $n_c$: | 4 | 8 | 7 | 9 | 28 |

**Figure 3** Example of Reliability Data in Canonical and Coincidence Matrix Terms

chance. The remaining values exhibit some agreement but also much uncertainty as to what the coded units are. The researcher would not know.

In the coincidence matrix, one might notice disagreements to follow a pattern. They occur exclusively near the diagonal of perfect agreements. There are no disagreements between extreme values, 1-4, or of the 1-3 and 2-4 kind. This pattern of disagreement would be expected in interval data. When the reliability data in Figure 3 are treated as interval data, $\alpha_{\text{interval}} = 0.877$. The interval $\alpha$ takes into account the proximity of the mismatching scale values—irrelevant in nominal data and appropriately ignored by $\alpha_{\text{nominal}}$. Hence, for data in Figure 3: $\alpha_{\text{nominal}} < \alpha_{\text{interval}}$. Had the disagreements been scattered randomly throughout the off-diagonal cells of the coincidence matrix, $\alpha_{\text{nominal}}$ and $\alpha_{\text{interval}}$ would not differ. Had disagreement been predominantly between the extreme values of the scale, e.g., 1-4, $\alpha_{\text{nominal}}$ would have exceeded $\alpha_{\text{interval}}$. This property of $\alpha$ provides the researcher with a diagnostic device to establish how coders use the given values.

## Statistical Properties of Alpha

A common mistake is to accept data as reliable when the null hypothesis that agreement results from chance fails. This test is seriously flawed as far as reliability is concerned. Reliable data need to contain no or only statistically insignificant disagreements. Acknowledging this requirement, a distribution of $\alpha$ offers two statistical indices: (1) *$\alpha$'s confidence limits,* ${}_{\text{low}}\alpha$ and ${}_{\text{high}}\alpha$, at a chosen level of statistical significance $p$, and more importantly, (2) the *probability $q$ that the measured $\alpha$ fails to exceed the ${}_{min}\alpha$ required* for data to be taken as sufficiently reliable. The distribution of $\alpha$ becomes narrower not only with increasing numbers of units sampled but also with increasing numbers of coders employed in the coding process.

The choice of ${}_{\text{min}}\alpha$ depends on the validity requirements of research undertaken with imperfect data. In academic research, it is common to aim for $\alpha \geq 0.9$ but require $\alpha \geq 0.8$ and accept data with $\alpha$ between 0.666 and 0.800 only to draw tentative conclusions. When human lives or valuable resources are at stake, ${}_{\text{min}}\alpha$ must be set higher.

To obtain the confidence limits for $\alpha$ at $p$ and probabilities $q$ for a chosen ${}_{\text{min}}\alpha$, the distributions of $\alpha_{\text{metric}}$ are obtained by bootstrapping in preference to mere mathematical approximations.

## Agreement Coefficients Embraced by Alpha

Alpha generalizes several known coefficients. It is defined to bring coefficients for different metrics but of the same makeup under the same roof. Alpha is applicable to any number of coders, which includes coefficients defined only for two. Alpha has no problem with missing data, as provided in Figure 1, which includes complete $m \times r$ data as a special case. Alpha corrects for small sample sizes, which includes the extreme of very large samples of data.

When data are nominal, generated by two coders, and consist of large sample sizes, $\alpha_{\text{nominal}}$ equals Scott's $\pi$. Scott's $\pi$, a popular and widely used coefficient in content analysis and survey research, conforms to $\alpha$'s conception of chance. When data are ordinal, generated by two coders, and very large, $\alpha_{\text{ordinal}}$ equals Spearmann's rank

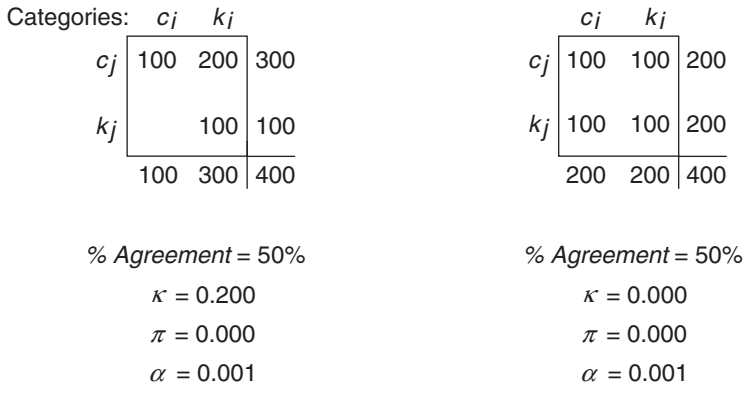



**Figure 4** Example of Kappa Adding Systematic Disagreements to the Reliability It Claims to Measure

correlation coefficient $\rho$ without ties in ranks. When data are interval, generated by two coders, and numerically large, $\alpha_{\text{interval}}$ equals Pearson's intraclass correlation coefficient $r_{ii}$. The intraclass correlation is the product moment correlation coefficient applied to symmetrical coincidence matrices rather than to asymmetrical contingency matrices. There is a generalization of Scott's $\pi$ to larger numbers of coders by Joseph Fleiss, who thought he was generalizing Cohen's $\kappa$, renamed $K$ by Sidney Siegel and John Castellan. $K$ equals $\alpha_{\text{nominal}}$ for a fixed number of coders with complete nominal data and very large, theoretically infinite sample sizes. Recently, there have been two close reinventions of $\alpha$, one by Kenneth Berry and Paul Mielke and one by Michael Fay.

## Agreement Coefficients Unsuitable as Indices of Data Reliability

Correlation coefficients for interval data and *association coefficients* for nominal data measure dependencies, statistical associations, between variables or coders, not agreements, and therefore cannot serve as measures of data reliability. In systems of correlations among many variables, the correlation among the same variables are often called reliabilities. They do not measure agreement, however, and cannot assess data reliability.

*Percent agreement*, limited to nominal data generated by two coders, varies from 0% to 100%, is the more difficult to achieve the more values are available for coding, and provides no indication about when reliability is absent. Percent agreement is not interpretable as a reliability scale—unless corrected for chance, which is what Scott's $\pi$ does.

*Cohen's 1960* κ (kappa), which also is limited to nominal data, two coders, and large sample sizes, has the undesirable property of counting systematic disagreement among coders as agreement. This is evident in unequal marginal distribution of categories in contingency matrices, which rewards coders who disagree on their use of categories with higher $\kappa$ values. Figure 4 shows two numerical examples of reliability data, tabulated in contingency matrices between two coders $i$ and $j$ in which terms $\kappa$ is originally defined.

Both examples show 50% agreement. They differ in their marginal distributions of categories. In the left example, data show coder $i$ to prefer category $k$ to category $c$ at a ratio of 3:1, whereas coder $j$ exhibits the opposite preference for $c$ over $k$ at the rate of 3:1—a systematic disagreement, absent in the data in the right example. The two examples have the same number of disagreements. Yet, the example with systematic disagreements measures $\kappa = 0.200$, whereas the one without that systematic disagreement measures $\kappa = 0.000$. In both examples, $\alpha = 0.001$. When sample sizes become large, $\alpha$ for two coders converges to Scott's $\pi$ at which point $\alpha = \pi = 0.000$. Evidently, Cohen's $\kappa$ gives "agreement credit" for this systematic disagreement, whereas $\pi$ and $\alpha$ do not. The reason for $\kappa$'s mistaken account of these systematic disagreements lies in Cohen's adoption of statistical independence between two coders as its conception of chance. This is customary when measuring

correlations or associations but has nothing to do with assigning units to categories. By contrast, in $\pi$ and $\alpha$, chance is defined as the statistical independence between the set of units coded and the values used to describe them. The margins of coincidence matrices estimate the distribution of values occurring in the population whose reliability is in question. The two marginal distributions of values in contingency matrices, by contrast, refer to coder preferences, not to population estimates. Notwithstanding its popularity, Cohen's $\kappa$ is inappropriate when the reliability of data is to be assessed.

Finally, *Cronbach's alpha* for interval data and *Kuder and Richardson's Formula-20* (KR-20) for binary data, which are widely used in psychometric and educational research, aim to measure the reliability of psychological tests by correlating the test results among multiple subjects. As Jum Nunnally and Ira Bernstein have observed, systematic errors are unimportant when studying individual differences. However, systematically biased coders reduce the reliability of the data they generate and such disagreement must not be ignored. Perhaps for this reason, Cronbach's alpha is increasingly interpreted as a measure on the internal consistency of tests. It is not interpretable as an index of the reliability of coded data.

*Klaus Krippendorff*

***See also*** Coefficient Alpha; Cohen's Kappa; Content Analysis; Interrater Reliability; KR-20; Replication; "Validity"

**Further Readings**

Berry, K. J., & Mielke, P. W., Jr. (1988). A generalization of Cohen's kappa agreement measure to interval measurement and multiple raters. *Educational and Psychological Measurement, 48,* 921–933.

Fay, M. P. (2005). Random marginal agreement coefficients: Rethinking the adjustments of chance when measuring agreement. *Biostatistics, 6,* 171–180.

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures, 1,* 77–89.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2nd ed.). Thousand Oaks, CA: Sage.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly, 19,* 321–325.

Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences* (2nd ed.). Boston: McGraw-Hill.

# Kruskal–Wallis Test

The Kruskal–Wallis test is a nonparametric test to decide whether $k$ independent samples are from different populations. Different samples almost always show variation regarding their sample values. This might be a result of chance (i.e., sampling error) if the samples are drawn from the same population, or it might be a result of a genuine population difference (e.g., as a result of a different treatment of the samples). Usually the decision between these alternatives is calculated by a one-way analysis of variance (ANOVA). But in cases where the conditions of an ANOVA are not fulfilled the Kruskal–Wallis test is an alternative approach because it is a nonparametric method; that is, it does not rely on the assumption that the data are drawn from a probability distribution (e.g., normal distribution).

Related nonparametric tests are the Mann–Whitney $U$ test for only $k = 2$ independent samples, the Wilcoxon signed rank test for $k = 2$ paired samples, and the Friedman test for $k > 2$ paired samples (repeated measurement) and are shown in Table 1.

The test is named after William H. Kruskal and W. Allen Wallis and was first published in the *Journal of the American Statistical Association* in 1952. Kruskal and Wallis termed the test as the $H$ test; sometimes the test is also named one-way analysis of variance by ranks.

**Table 1** Nonparametric Tests to Decide Whether $k$ Samples Are From Different Populations

| | $k = 2$ | $k > 2$ |
|---|---|---|
| *Independent samples* | Mann–Whitney $U$ test | Kruskal–Wallis test |
| *Paired samples* | Wilcoxon signed rank test | Friedman test |

**Table 2** Example Data, $N = 15$ Observations From $k = 3$ Different Samples (Groups)

| *Group 1* | | *Group 2* | | *Group 3* | |
|---|---|---|---|---|---|
| *Observation* | *Rank* | *Observation* | *Rank* | *Observation* | *Rank* |
| 2400 | 14 | 1280 | 4.5 | 1280 | 4.5 |
| 1860 | 10 | 1690 | 9 | 1090 | 1 |
| 2240 | 13 | 1890 | 11 | 2220 | 12 |
| 1310 | 7 | 1100 | 2 | 1310 | 7 |
| 2700 | 15 | 1210 | 3 | 1310 | 7 |
| $R_1$ | 59 | $R_2$ | 29.5 | $R_3$ | 31.5 |

*Notes:* Next to the observation, the individual rank for this observation can be seen. Ranks are added up per sample to a rank sum termed $R_i$.

This entry begins with a discussion of the concept of the Kruskal–Wallis test and provides an example. Next, this entry discusses the formal procedure and corrections for ties. Last, this entry describes the underlying assumptions of the Kruskal–Wallis test.

## Concept and Example

The idea of the test is to bring all observations of all $k$ samples into a rank order and to assign them an according rank. After this initial procedure, all further calculations are based only on these ranks but not on the original observations anymore. The underlying concept of the test is that these ranks should be equally distributed throughout the $k$ samples, if all observations are from the same population. A simple example is used to demonstrate this.

A researcher made measurements on $k = 3$ different groups. Overall there are $N = 15$ observations. Data are arranged according to their group, and an individual rank is assigned to each observation starting with 1 for the smallest observation (see Table 2).

Two things might be noted here. First, in this example, for the sake of simplicity, all groups have the same number of observations; however, this is not a necessary condition. Second, there are several observations with the same value called *tie*. In this case, all observations sharing the same values are assigned their mean rank. In the current example, two observations resulted in a value of 1280 sharing ranks 4 and 5. Thus, both receive rank 4.5. Furthermore, three samples had a value of 1310 and would have received ranks from 6 to 8. Now they get rank 7 as the mean of 6, 7, and 8.

In the next step, the sum of ranks $(R_1, R_2, R_3)$ for each group is calculated. The overall sum of ranks is $N(N+1)/2$. In the example case, this is $15 \times 16 / 2 = 120$. As a first control, the sum of ranks for all groups should add up to the same value: $R_1 + R_2 + R_3 = 59 + 29.5 + 31.5 = 120$.

Distributing the ranks among the three groups randomly, each rank sum would be about $120/3 = 40$. The idea is to measure and add the squared deviations from this expectancy value:

$$(59 - 40)^2 + (29.5 - 40)^2 + (31.5 - 40)^2 = 543.5.$$

The test statistic provided by the formula of Kruskal and Wallis is a transformation of this sum of squares, and the resulting $H$ is under certain conditions (see below) chi-square distributed for $k - 1$ degrees of freedom. In the example case, the conditions for an asymptotic chi-square test are not fulfilled because there are not enough observations. Thus, the probability for $H$ has to be looked up in a table, which can be found in many statistics books and on the Internet. In this case, we calculate $H = 5.435$ according to the formula provided in the following section. The critical value for this case ($k = 3, 5$ observations in each group) for $p = .05$ can be found in the table; it is $H = 5.780$. Thus, for the example shown earlier, the null hypothesis that all three samples are from the same population will not be rejected.

## Formal Procedure

The Kruskal–Wallis test assesses the null hypothesis that $k$ independent random samples are from the same population. Or to state it more precisely that the samples come from populations with the same locations. Because the logic of the test is based on comparing ranks rather than means or medians, it is not correct to say that it tests the equality of means or medians in populations.

Overall $N$ observations are made in $k$ samples. All observations have to be brought into a rank order and ranks have to be assigned from the smallest to the largest observation. No attention is given to the sample to which the observation belongs. In

the case of shared ranks (ties), the according mean rank value has to be assigned. Next a sum of ranks $R_i$ for all $k$ samples (from $i = 1$ to $i = k$) has to be computed. The according number of observations of each sample is denoted by $N_i$, the number of all observations by $N$. The Kruskal–Wallis test statistic $H$ is computed according to

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{N_i} - 3(N+1).$$

For larger samples, $H$ is approximately chi-square distributed with $k - 1$ degrees of freedom. For smaller samples, an exact test has to be performed and the test statistic $H$ has to be compared with critical values in tables, which can be found in statistics books and on the Internet. (The tables provided by Kruskal and Wallis in the *Journal of the American Statistical Association,* 1952, 47, 614–617, contain some errors; an errata can be found in the *Journal of the American Statistical Association,* 1953, 48, 910.) These tables are based on a full permutation of all possible rank distributions for a certain case. By this technique, an empirical distribution of the test criterion $H$ is obtained by a full permutation. Next the position of the obtained $H$ within this distribution can be determined. The according $p$ value reflects the cumulative probability of $H$ to obtain this or even a larger value by chance alone. There is no consistent opinion what exactly forms a small sample. Most authors recommend for $k = 3$ and $N_i \leq 8$ observations per sample, for $k = 4$ and $N_i \leq 4$, and for $k = 5$ and $N_i \leq 3$ to perform the exact test.

## Correction for Ties

When ties (i.e., shared ranks) are involved in the data, there is a possibility of correcting for this fact when computing $H$.

$$H_{\text{corr}} = \frac{H}{C};$$

thereby $C$ is computed by

$$C = 1 - \frac{\sum_{i=1}^{m} (t_i^3 - t_i)}{N^3 - N}.$$

In this formula, $m$ stands for the number of ties occurred, and $t_i$ stands for the number of tied ranks occurred in a specific tie $i$.

In the preceding example, there were $m = 2$ tied observations, one for the ranks 4 to 5, and another one for the ranks ranging from 6 to 8. For the first tie $t_1 = 2$, since two observations are tied, for the second tie $t_2 = 3$, because three observations are identical. Thus, we compute the following correction coefficient:

$$C = 1 - \frac{(2^3 - 2) + (3^3 - 3)}{15^3 - 15} = 0.991;$$

with this coefficient, $H_{\text{corr}}$ calculates to

$$H_{\text{corr}} = \frac{5.435}{0.991} = 5.484.$$

Several issues might be noted as seen in this example. The correction coefficient will always be smaller than one, and thus, $H$ will always increase by this correction formula. If the null hypothesis is already rejected by an uncorrected $H$, any further correction might strengthen the significance of the result, but it will never result in not rejecting the null hypothesis. Furthermore, the correction of $H$ resulting from this computation is very small, even though 5 out of 15 (33%) of the observations in the example were tied. Even in this case where the uncorrected $H$ was very close to significance, the correction was negligible. From this perspective, it is only necessary to apply this correction, if $N$ is very small or if the number of tied observations is relatively large compared with $N$; some authors recommend here a ratio of 25%.

## Assumptions

The Kruskal–Wallis test does not assume a normal distribution of the data. Thus, whenever the requirement for a one-way ANOVA to have normally distributed data is not met, the Kruskal–Wallis test can be applied instead. Compared with the $F$ test, the Kruskal–Wallis test is reported to have an asymptotic efficiency of 95.5%.

However, several other assumptions have to be met for the Kruskal–Wallis test:

1. Variables must have at least an ordinal level (i.e., rank-ordered data).

2. All samples have to be random samples, and the samples have to be mutually independent.
3. Variables need to be continuous, although a moderate number of ties is tolerable as shown previously.
4. The populations from which the different samples drawn should only differ regarding their central tendencies but not regarding their overall shape. This means that the populations might differ, for example, in their medians or means, but not in their dispersions or distributional shape (such as, e.g., skewness). If this assumption is violated by populations of dramatically different shapes, the test loses its consistency (i.e., a rejection of the null hypothesis by increasing $N$ is not anymore guaranteed in a case where the null hypothesis is not valid).

*Stefan Schmidt*

***See also*** Analysis of Variance (ANOVA); Friedman Test; Mann–Whitney *U* Test; Nonparametric Statistics; Wilcoxon Rank Sum Test

**Further Readings**

Conover, W. J. (1971). *Practical nonparametric statistics*. New York: Wiley.
Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association, 47,* 583–621.
Kruskal, W. H., & Wallis, W. A. (1953). Errata: use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association, 48,* 907–911.

# Kurtosis

Density functions are used to describe the distribution of quantitative variables. Kurtosis is a characteristic of the shape of the density function related to both the center and the tails. Distributions with density functions that have significantly more mass toward the center and in the tails than the normal distribution are said to have high kurtosis. Kurtosis is invariant under changes in location and scale; thus, kurtosis remains the same after a change in units or the standardization of data.

There are several alternative ways of measuring kurtosis; they differ in their sensitivity to the tails of the distribution and to the presence of outliers.

Some tests of normality are based on the comparison of the skewness and kurtosis of the data with the values corresponding to a normal distribution. Tools to do inference about means and variances, many of them developed under the assumption of normality, see their performance affected when applied to data from a distribution with high kurtosis.

The next two sections focus on kurtosis of theoretical distributions, and the last two deal with kurtosis in the data analysis context.

## Comparing Distributions in Terms of Kurtosis

A distribution such as the Laplace is said to have higher kurtosis than the normal distribution because it has more mass toward the center and heavier tails [see Figure 1(a)]. To visually compare the density functions of two symmetric distributions in terms of kurtosis, these should have the same center and variance. Figure 1(b) displays the corresponding cumulative version or distribution functions (CDFs). Willem R. van Zwet defined in 1964 a criterion to compare and order symmetric distributions based on their CDFs. According to this criterion, the normal has indeed no larger kurtosis than the Laplace distribution. However, not all the symmetric distributions are ordered.

## Measuring Kurtosis in Distributions

In Greek, *kurtos* means convex; the mathematician Heron in the first century used the word *kurtosis* to mean curvature. Kurtosis was defined, as a statistical term, by Karl Pearson around 1905 as the measure

$$\beta_2 = E(x-\mu)^4/\sigma^4$$

to compare other distributions with the normal distribution in terms of the frequency toward the mean $\mu$ ($\sigma$ is the standard deviation and $\beta_2 = 3$ for the normal distribution). It was later that ordering criteria based on the distribution functions were defined; in addition, more flexible definitions acknowledging that kurtosis is related

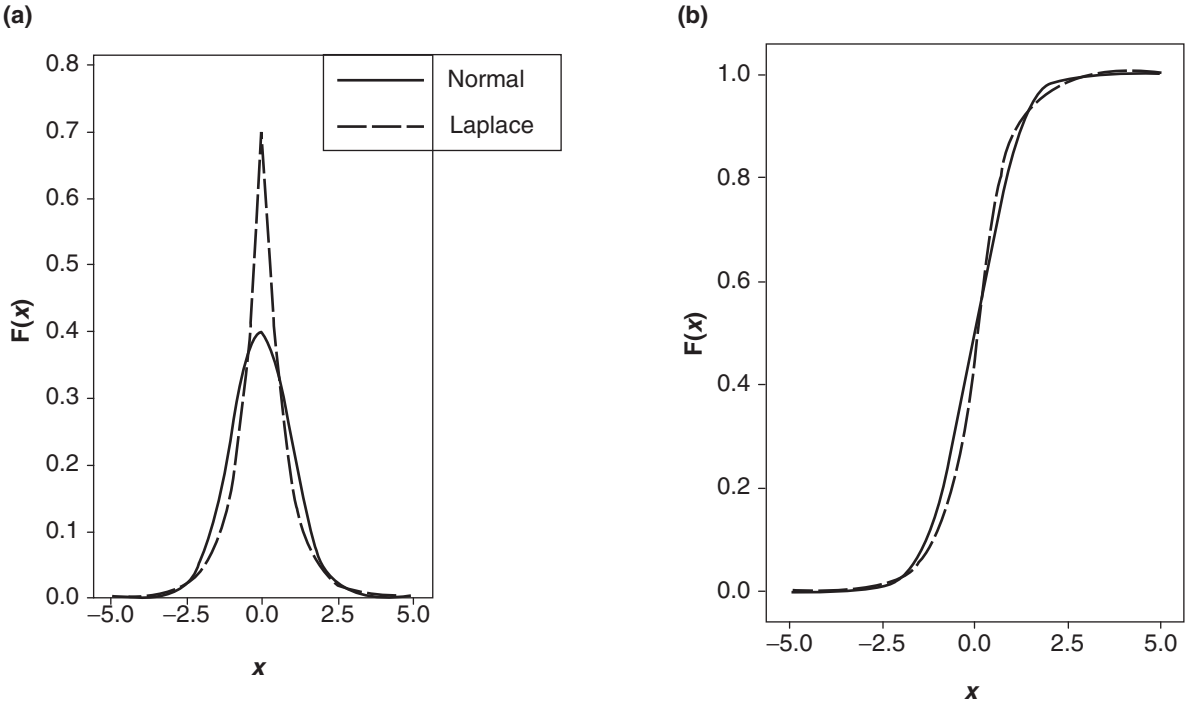



**Figure 1** Density Functions and CDFs for Normal and Laplace Distributions
*Notes:* (a) Density functions. (b) Cumulative distribution functions.

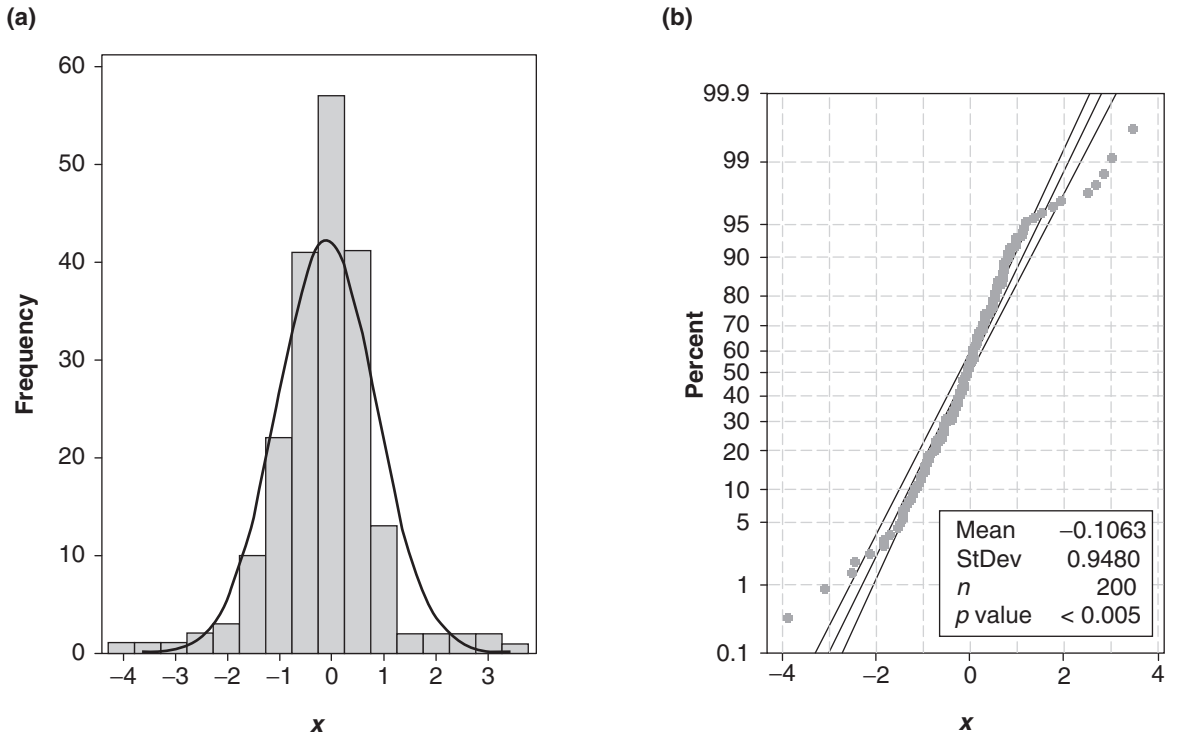



**Figure 2** Histogram and Normal Probability Plot for a Sample From a Laplace Distribution
*Notes:* (a) Histogram of data. (b) Probability plot of data. Normal = –95% CI.

**Table 1** Values of Some Kurtosis Measures for Eight Symmetric Distributions

| *Distribution* | $\beta_2$ | $\gamma_2(0.05)$ | $\tau_4$ | *Distribution* | $\beta_2$ | $\gamma_2(0.05)$ | $\tau_4$ |
|---|---|---|---|---|---|---|---|
| Uniform | 1.8 | 0 | 0 | $t_{(4)}$ | | 0.503 | 0.217 |
| Normal | 3 | 0.355 | 0.123 | $t_{(2)}$ | | 0.648 | 0.375 |
| Laplace | 6 | 0.564 | 0.236 | $t_{(1)}$ | | 0.854 | 1 |
| SU(0,1) | 36.2 | 0.611 | 0.293 | SU(0,0.9) | 82.1 | 0.649 | 0.329 |

to both peakedness and tail weight were proposed, and it was accepted that kurtosis could be measured in several ways. New measures of kurtosis, to be considered valid, have to agree with the orderings defined over distributions by the criteria based on distribution functions. It is said that some kurtosis measures, such as $\beta_2$, naturally have an averaging effect that prevents them from being as informative as the CDFs. Two distributions can have the same value of $\beta_2$ and still look different. Because kurtosis is related to the peak and tails of a distribution, in the case of nonsymmetric distributions, kurtosis and skewness tend to be associated, particularly if they are represented by measures that are highly sensitive to the tails.

Two of the several kurtosis measures that have been defined as alternatives to $\beta_2$ are as follows:

1. L-kurtosis defined by J. R. M. Hosking in 1990 and widely used in the field of hydrology; $\tau_4 = L_4/L_2$ is a ratio of $L-$*moments* that are linear combinations of expected values of order statistics.
2. Quantile kurtosis, $\gamma_2(p)$, defined by Richard Groeneveld in 1998 for symmetric distributions only, is based on distances between certain quantiles. Other kurtosis measures defined in terms of quantiles, quartiles, and octiles also exist.

As an illustration, the values of these kurtosis measures are displayed in Table 1 for some distributions.

## Studying Kurtosis From Data

Histograms and probability plots help to explore sample data. Figure 2 indicates that the data might come from a distribution with higher kurtosis than the normal. Statistical software generally calculates the excess $\hat{\beta}_2 - 3$ (for the data in Figure 2, $\hat{\beta}_2 - 3 = 3.09$).

If the sample size $n$ is small,

$$b_2 = \frac{\left[\sum (x - \bar{x})^4\right]/n}{s^4}$$

might take a small value even if the kurtosis of the population is very high [the upper bound for $b_2$ is $n - 2 + 1/(n-1)$]. Adjusted estimators exist to reduce the bias, at least in the case of nearly normal distributions. A commonly used adjusted estimator of excess is

$$\frac{n(n+1)}{(n-1)(n-2)(n-3)} \frac{\left[\sum (x - \bar{x})^4\right]}{s^4} - \frac{3(n-1)(n-1)}{(n-2)(n-3)}.$$

A single distant outlier can dramatically change the value of $\hat{\beta}_2$.

## Effects of High Kurtosis

The variance of the sample variance is related to $\beta_2$. The power of some tests for the equality of variances gets affected by high kurtosis. For example, when testing the hypothesis of equal variances for two populations based on two independent samples, the power of the Levene test is lower if the distribution of the populations is symmetric but with higher kurtosis than if the samples come from normal distributions. The performance of the $t$ test for the population mean is also affected under situations of high kurtosis. Van Zwet proved that, when working with symmetric distributions, the median is more efficient than the mean as estimator of the center of the distribution if the latter has very high kurtosis.

*Edith Seier*

*See also* Distribution; Median; Normal Distribution; Student's *t* Test; Variance

## Further Readings

Balanda, K. P., & MacGillivray, H. L. (1998). Kurtosis: A critical review. *American Statistician, 42,* 111–119.

Bonett, D. G., & Seier, E. (2002). A test of normality with high uniform power. *Computational Statistics and Data Analysis, 40,* 435–445.

Byers, R. H. (2000). On the maximum of the standardized fourth moment. *InterStat,* January.

Hosking, J. R. M. (1992). Moments or L moments? An example comparing two measures of distributional shape. *American Statistician, 46*, 186–199.

Ruppert, D. (1997). What is kurtosis? An influence function approach. *American Statistician, 41,* 1–5.

Seier, E., & Bonett, D. G. (2003). Two families of kurtosis measures. *Metrika, 58,* 59–70.

# L

## L'Abbé Plot

The L'Abbé plot is one of several graphs commonly used to display data visually in a meta-analysis of clinical trials that compare a treatment and a control intervention. It is basically a scatterplot of results of individual studies with the risk in the treatment group on the vertical axis and the risk in the control group on the horizontal axis. This plot was advocated in 1987 by Kristan L'Abbé and colleagues for visually showing variations in observed results across individual trials in meta-analysis. This entry briefly discusses meta-analysis before addressing the usefulness, limitations, and inappropriate uses of the L'Abbé plot.

### Meta-Analysis

To understand what the L'Abbé plot is, it is necessary to have a discussion about meta-analysis. Briefly, meta-analysis is a statistical method to provide a summary estimate by combining the results of many similar studies. A hypothesized meta-analysis of 10 clinical trials is used here to illustrate the use of the L'Abbé plot. The most commonly used graph in meta-analysis is the *forest plot* (as shown in Figure 1) to display data from individual trials and the summary estimate (including point estimates and 95% confidence intervals). The precision or statistical power of the summary estimate will be much improved by combining the results of many small studies. In this hypothesized meta-analysis, the pooled estimate of relative risk is 0.72 (95% confidence interval: 0.53–0.97), which suggests that the risk of lack of clinical improvement in the treatment group is statistically significantly lower than that in the control group. However, the results from the 10 trials vary considerably (Figure 1), and it is important to investigate why similar trials of the same intervention might yield different results.

Figure 2 is the L'Abbé plot for the hypothesized meta-analysis. The vertical axis shows the event rate (or risk) of a lack of clinical improvement in the treatment group, and the horizontal axis shows the event rate of a lack of clinical improvement in the control group. Each point represents the result of a trial, according to the corresponding event rates in the treatment and the control group. The size of the points is proportionate to the trial size or the precision of the result. The larger the sample size, the larger the point in Figure 2. However, it should be mentioned that smaller points might represent larger trials in a L'Abbé plot produced by some meta-analysis software.

The diagonal line (line A) in Figure 2 is called the *equal line,* indicating the same event rate between the two arms within a trial. That is, a trial point will lie on the equal line when the event rate in the treatment group equals that in the control group. Points below the equal line indicate that the risk in the treatment group is lower than that

| Study or Subgroup | Treatment Events | Treatment Total | Control Events | Control Total | Weight | Risk Ratio IV, Random, 95% CI |
|---|---|---|---|---|---|---|
| Trial-01 | 20 | 400 | 16 | 400 | 10.6% | 1.25 [0.66, 2.38] |
| Trial-02 | 58 | 800 | 60 | 800 | 15.9% | 0.97 [0.68, 1.37] |
| Trial-03 | 11 | 100 | 12 | 100 | 8.8% | 0.92 [0.42, 1.98] |
| Trial-04 | 9 | 50 | 7 | 50 | 7.2% | 1.29 [0.52, 3.18] |
| Trial-05 | 26 | 200 | 36 | 200 | 13.7% | 0.72 [0.45, 1.15] |
| Trial-06 | 14 | 100 | 20 | 100 | 10.9% | 0.70 [0.37, 1.31] |
| Trial-07 | 6 | 50 | 14 | 50 | 7.6% | 0.43 [0.18, 1.03] |
| Trial-08 | 7 | 25 | 8 | 25 | 7.8% | 0.88 [0.37, 2.05] |
| Trial-09 | 14 | 100 | 41 | 100 | 12.3% | 0.34 [0.20, 0.59] |
| Trial-10 | 3 | 25 | 12 | 25 | 5.3% | 0.25 [0.08, 0.78] |
| **Total (95% CI)** | | **1850** | | **1850** | **100.0%** | **0.72 [0.53, 0.97]** |
| Total events | 168 | | 226 | | | |

Heterogeneity: Tau² = 0.12; Chi² = 19.63, df = 9 ($p$ = .02); I² = 54%

Test for overall effect: Z = 2.16 ($p$ = .03)

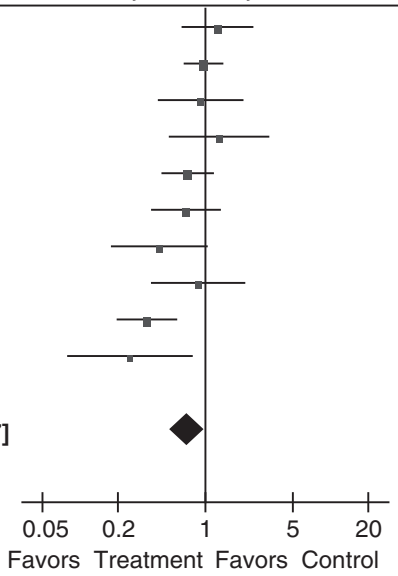
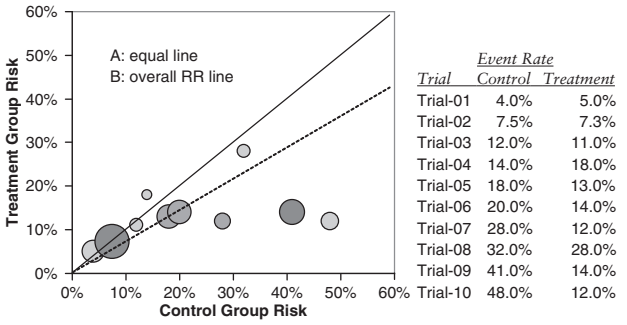



**Figure 1** A Hypothesized Meta-Analysis of 10 Clinical Trials Comparing a Treatment and a Control Intervention

*Note:* Outcome: lack of clinical improvement.




| Trial | Event Rate Control | Event Rate Treatment |
|---|---|---|
| Trial-01 | 4.0% | 5.0% |
| Trial-02 | 7.5% | 7.3% |
| Trial-03 | 12.0% | 11.0% |
| Trial-04 | 14.0% | 18.0% |
| Trial-05 | 18.0% | 13.0% |
| Trial-06 | 20.0% | 14.0% |
| Trial-07 | 28.0% | 12.0% |
| Trial-08 | 32.0% | 28.0% |
| Trial-09 | 41.0% | 14.0% |
| Trial-10 | 48.0% | 12.0% |

**Figure 2** The L'Abbé Plot for the Hypothesized Meta-Analysis of 10 Clinical Trials

in the control group, and vice versa for points above the equal line. In the hypothesized meta-analysis, the central points of two trials (T01 and T04) are above the equal line, indicating that the event rate in the treatment group is higher than that in the control group. The points of the remaining eight trials locate below the equal line, showing that the risk in the treatment group is reduced in these trials.

The dotted line (line B) in Figure 2 is the *overall RR line*, which represents the pooled relative risk in meta-analysis. This overall RR line corresponds to a pooled relative risk of 0.72 in Figure 2. It would be expected that the points of most trials will lie around this overall RR line. The distance between a trial point and the overall RR line indicates the difference between the trial result and the average estimate.

## Usefulness

Clinical trials that evaluate the same underlying average treatment effect might generate different results because of random error. The smaller the sample size, the greater will be the random variation. Meta-analysis can yield a weighted average by pooling the results of all available similar trials. This weighted average is the best summary estimate of the true treatment effect if the variation in results across trials is mainly caused by random error. However, the variation in results across trials often cannot be explained satisfactorily by chance alone. For example, the result of statistical testing of heterogeneity in the hypothesized meta-analysis (Figure 1) suggests that the heterogeneity across trials is statistically significant ($p = .02$). Different

patient and/or intervention characteristics might also be the causes of variation in results across trials. The effect of a treatment might be associated with the severity of illness, age, gender, or other patient characteristics. Trial results might vary because of different doses of medications, different intensity of interventions, different level of training and experience of doctors, and other differences in settings or interventions.

The variation in results across studies is termed *heterogeneity* in the meta-analysis. Several graphical methods can be used for the investigation of heterogeneity in meta-analysis. The commonly used graphical methods are forest plot, funnel plot, Galbraith plot, and the L'Abbé plot. Only estimates of relative effects (including relative risk, odds ratio, or risk difference) between the treatment and control group are displayed in forest plot, funnel plot, and Galbraith plot. As compared with other graphical methods, an advantage with the L'Abbé plot in meta-analysis is that it can reveal not only the variations in estimated relative effects across individual studies but also the trial arms that are responsible for such differences. This advantage of the L'Abbé plot might help researchers and clinicians to identify the focus of the investigation of heterogeneity in meta-analysis.

In the hypothesized meta-analysis (Figure 2), the event rate varies greatly in both the control group (from 4.0% to 48.0%) and in the treatment group (from 5.0% to 28.0%). The points of trials with relatively low event rates in the control group tend to locate above the overall RR line, and the points of trials with relatively high event rates in the control group tend to locate below the overall RR line. This suggests that variations in relative risk across trials might be mainly a result of different event rates in the control group. Therefore, the event rate in the control group might be associated with treatment effect in meta-analysis. In a real meta-analysis, this pattern of graphical distribution should be interpreted by considering other patient and/or intervention characteristics to investigate the possible causes of variations in results across trials.

## Limitations

A shortcoming of many graphical methods is that the visual interpretation of data is subjective, and the same plot might be interpreted differently by different people. In addition, the same pattern of variations across studies revealed in a L'Abbé plot might have very different causes. The usefulness of the L'Abbé plot is also restricted by the available data reported in the primary studies in meta-analysis. When the number of the available studies in a meta-analysis is small and when data on important variables are not reported, the investigation of heterogeneity will be unlikely fruitful.

The visual perception of variations across studies in a L'Abbé plot might be misleading because random variation in the distance between a study point and the overall RR line is associated with both the sample size of a trial and the event rate in the control group. Points of small trials are more likely farther away from the overall RR line purely by chance. In addition, trials with a control event rate closing to 50% will have great random variation in the distance from the overall RR line. It is possible that the distances between trial points and the overall RR line in a L'Abbé plot are adjusted by the corresponding sample sizes and event rates in the control group, using a stochastic simulation approach. However, the stochastic simulation method is complex and cannot be used routinely in meta-analysis.

## Inappropriate Uses

L'Abbé plots have been used in some meta-analyses to identify visually the trial results that are outliers according to the distance between a trial point and the overall RR line. Then, the identified outliers are excluded from meta-analysis one by one until heterogeneity across studies is no longer statistically significant. However, this use of the L'Abbé plot is inappropriate because of the following reasons. First, the exclusion of studies according to their results, not their design and other study characteristics, might introduce bias into meta-analysis and reduce the power of statistical tests of heterogeneity in meta-analysis. Second, the chance of revealing clinically important causes of heterogeneity might be missed simply by excluding studies from meta-analysis without efforts to investigate reasons for the observed heterogeneity. In addition, different methods might identify different trials as outliers, and the exclusion of different studies might lead to different results of the same meta-analysis.

Another inappropriate use of the L'Abbé plot is to conduct a regression analysis of the event rate

in the treatment group against the event rate in the control group. If the result of such a regression analysis is used to examine the relation between treatment effect and the event rate in the control group, then misleading conclusions could be obtained. This is because of the problem of regression to the mean and random error in the estimated event rates.

### Final Thoughts

To give an appropriate graphical presentation of variations in results across trials in meta-analysis, the scales used on the vertical axis and on the horizontal axis should be identical in a L'Abbé plot. The points of trials should correspond to the different sample sizes or other measures of precision. It should be explicit about whether the larger points correspond to larger or smaller trials. As compared with other graphs in meta-analysis, one important advantage of the L'Abbé plot is that it displays not only the relative treatment effects of individual trials but also data on each of the two trial arms. Consequently, the L'Abbé plot is helpful to identify not only the studies with outlying results but also the study arm being responsible for such differences. Therefore, the L'Abbé plot is a useful graphical method for the investigation of heterogeneity in meta-analysis. However, the interpretation of L'Abbé plots is subjective. Misleading conclusions could be obtained if the interpretation of a L'Abbé plot is inappropriate.

*Fujian Song*

***See also*** Bias; Clinical Trial; Control Group; Meta-Analysis; Outlier; Random Error

#### Further Readings

Bax, L., Ikeda, N., Fukui, N., Yaju, Y., Tsuruta, H., & Moons, K. G. M. (2008). More than numbers: The power of graphs in meta-analysis. *American Journal of Epidemiology, 169,* 249–255.

L'Abbé, K. A., Detsky, A. S., & O'Rourke, K. (1987). Meta-analysis in clinical research. *Annals of Internal Medicine, 107,* 224–233.

Sharp, S. J., Thompson, S. G., & Altman, D. G. (1996). The relation between treatment benefit and underlying risk in meta-analysis. *British Medical Journal, 313,* 735–738.

Song, F. (1999). Exploring heterogeneity in meta-analysis: Is the L'Abbé plot useful? *Journal of Clinical Epidemiology, 52,* 725–730.

Song, F., Sheldon, T. A., Sutton, A. J., Abrams, K. R., & Jones, D. R. (2001). Methods for exploring heterogeneity in meta-analysis. *Evaluation and the Health Professions, 24,* 126–151.

## Laboratory Experiments

Laboratory experiments are a particular method that enables the highest level of control for hypothesis testing. Like other types of experiments, they use random assignment and intentional manipulations, but these experiments are conducted in a room or a suite of rooms dedicated to that purpose. Although experimental research can be conducted in places besides laboratories, such as in classrooms or business organizations, a laboratory setting is usually preferable, because an investigator can create optimal conditions for testing the ideas guiding the research.

Psychology was the first social science to use experimental laboratories, with Ivan Pavlov's famous experiments conditioning dogs around the turn of the 20th century. However, it was the second half of that century that saw the spread of experiments throughout the other social sciences. In the 1940s and 1950s, R. Freed Bales conducted discussion groups at Harvard, developing a research design still used for many purposes, including focus groups in communications studies and marketing.

Bales's groups, for reasons including practical limitations, included no more than about 20 individuals, and experimental studies were once called "small group" research. A more accurate term would be "group processes" research, because the focus of study is not actually the group but rather what happens within the group. Researchers do not so much study the group itself as they study abstract processes that occur in interaction. Experimenters cannot study an army or a business corporation in the laboratory, but they can and do study authority structures, negotiation processes, responses to legitimate or illegitimate orders, status generalization, and communication within networks. In other words, abstract features of

concrete structures such as the U.S. Army can be studied experimentally. The results of laboratory research, with proper interpretation, can then be applied in businesses, armies, or other situations meeting the conditions of the theory being tested by the experiment.

## Laboratories as Created Situations

The essential character of a laboratory experiment is that it creates an invented social situation that isolates theoretically important processes. Such a situation is usually unlike any naturally occurring situation so that complicated relationships can be disentangled. In an experiment, team partners might have to resolve many disagreements over a collective task, or they might be asked to decide whether to offer a small gift to people who always (or never) reciprocate. In such cases, an investigator is efficiently studying things that occur naturally only occasionally, or that are hard to observe in the complexity of normal social interaction.

Bringing research into a laboratory allows an investigator to simplify the complexity of social interaction to focus on the effects of one or a few social processes at a time. It also offers an opportunity to improve data collection greatly using video and sound recordings, introduce questionnaires at various points, and interview participants about their interpretations of the situation and of people's behavior in it.

Every element of the social structure, the interaction conditions, and the independent variables is included in the laboratory conditions because an investigator put it there. The same is true for the measurement operations used for the dependent variables. Well-designed experiments result from a thorough understanding of the theoretical principles to be tested, long-term planning, and careful attention to detail. Casually designed experiments often produce results that are difficult to interpret, either because it is not clear exactly what happened in them or because measurements seem to be affected by unanticipated, perhaps inconsistent, factors.

## Strong Tests and Inferences

All laboratories simplify nature. Just as chemistry laboratories contain pure chemicals existing nowhere in nature, social science laboratories contain situations isolating one or a few social processes for detailed study. The gain from this focus is that experimental results are among the strongest for testing hypotheses.

Most hypotheses, whether derived from general theoretical principles or simply formulated ad hoc, have the form "If A then B." To test such a sentence, treat "A" as an independent variable and "B" as a dependent variable. Finding that B is present when A is also present gives some confidence the hypothesis is correct, but of course the concern is that something else besides A better accounts for the presence of B. But when the copresence of A and B occur in a laboratory, an investigator has had the opportunity to remove other possible candidates besides A from the situation. In a laboratory test, the experimental situation creates A and then measures to determine the existence or the extent of B.

Another concern in natural settings is direction of causality. Finding A and B together is consistent with (a) A causes B; (b) B causes A; or (c) some other factor C causes both A and B.

Although we cannot ever observe causation, even with laboratory data, such data can lend greater confidence in hypothesis (a). That is because an experimenter can introduce A before B occurs, thus making interpretation (b) unlikely. An experimenter can also simplify the laboratory situation to eliminate plausible Cs, even if some unknown factor might still be present that is affecting B. In general, the results from laboratory experiments help in assessing the directionality of causation and eliminating potential alternative explanations of observed outcomes.

## Laboratories Abstract From Reality

Experimental research requires conceptualizing problems abstractly and generally. Many important questions in social science are concrete or unique, and therefore these questions do not lend themselves readily to laboratory experimentation. For instance, the number of homeless in the United States in a particular year is not a question for laboratory methods. However, effects of network structures, altruism, and motivational processes that might affect homelessness

are suitable for experimental study. Abstract and general conceptualization loses many concrete features of unique situations while gaining applicability to other situations besides the one that initially sparked interest. From laboratory studies, we will never know how many Americans are homeless, but we might learn that creating social networks of particular kinds is a good way to reduce that number.

### Ethics

Ethical considerations are crucial in all social science research. Unfortunately, some infamous cases in biomedical and even in social research have sometimes caused observers to associate ethical malfeasance and laboratory experiments. Such linkage is unwarranted. Protecting participants' rights and privacy are essential parts of any sort of research. Institutional Review Boards (IRBs) oversee the protection of human research participants and the general rules of justice, beneficence, and respect guide the treatment of experimental participants. Experimental participants are volunteers who receive incentives (e.g., money or course credit) for their work, and good research design includes full explanations and beneficial learning experiences for participants. There is no place for neglect or mistreatment of participants.

*Murray Webster, Jr., and Jane Sell*

***See also*** Ethics in the Research Process; Experimental Design; Experimenter Expectancy Effect; Quasi-Experimental Design

#### Further Readings

Rashotte, L. S., Webster, M., & Whitmeyer, J. (2005). Pretesting experimental instructions. *Sociological Methodology, 35,* 151–175.

Webster, M., & Sell, J. (Eds.). (2007). *Laboratory experiments in the social sciences.* New York: Elsevier.

Webster, M., & Sell, J. (2006). Theory and experimentation in the social sciences. In W. Outhwaite & S. P. Turner (Eds.), *The Sage handbook of social science methodology* (pp. 190–207). Thousand Oaks, CA: Sage.

## Last Observation Carried Forward

Last observation carried forward (LOCF) is a method of imputing missing data in longitudinal studies. If a person drops out of a study before it ends, then his or her last observed score on the dependent variable is used for all subsequent (i.e., missing) observation points. LOCF is used to maintain the sample size and to reduce the bias caused by the attrition of participants in a study. This entry examines the rationale for, problems associated with, and alternative to LOCF.

### Rationale

When participants drop out of longitudinal studies (i.e., ones that collect data at two or more time points), two different problems are introduced. First, the sample size of the study is reduced, which might decrease the power of the study, that is, its ability to detect a difference between groups when one actually exists. This problem is relatively easy to overcome by initially enrolling more participants than are actually needed to achieve a desired level of power, although this might result in extra cost and time. The second problem is a more serious one, and it is predicated on the belief that people do not drop out of studies for trivial reasons. Patients in trials of a therapy might stop coming for return visits if they feel they have improved and do not recognize any further benefit to themselves from continuing their participation. More often, however, participants drop out because they do not experience any improvement in their condition, or they find the side effects of the treatment to be more troubling than they are willing to tolerate. At the extreme, the patients might not be available because they have died, either because their condition worsened or, in rare cases, because the "treatment" actually proved to be fatal. Thus, those who remain in the trial and whose data are analyzed at the end reflect a biased subset of all those who were enrolled. Compounding the difficulty, the participants might drop out of the experimental and comparison groups at different rates, which biases the results even more.

Needless to say, the longer the trial and the more follow-up visits or interviews that are required, the worse the problem of attrition becomes. In some clinical trials, drop-out rates approach 50% of those who began the study.

LOCF is a method of data imputation, or "filling in the blanks," for data that are missing because of attrition. This allows the data for all participants to be used, ostensibly solving the two problems of reduced sample size and biased results. The method is quite simple, and consists of replacing all missing values of the dependent variable with the last value that was recorded for that particular participant. The justification for using this technique is shown in Figure 1, where the left axis represents symptoms, and lower scores are better. If the effect of the treatment is to reduce symptoms, then LOCF assumes that the person will not improve any more after dropping out of the trial. Indeed, if the person discontinues very early, then there might not be any improvement noted at all. This most probably underestimates the actual degree of improvement experienced by the patient and, thus, is a conservative bias; that is, it works against the hypothesis that the intervention works. If the findings of the study are that the treatment does work, then the researcher can be even more confident of the results. The same logic applies if the goal of treatment is to increase the score on some scale; LOCF carries forward a smaller improvement.

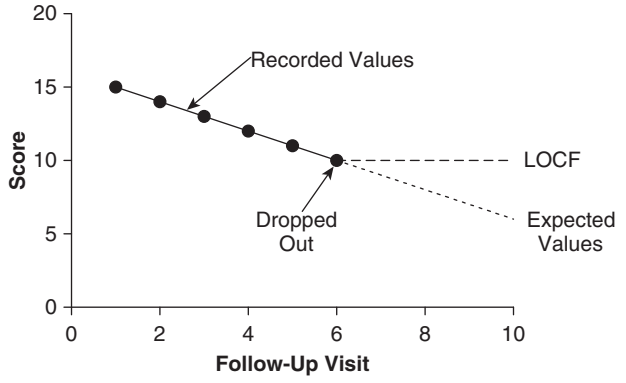



**Figure 1** The Rationale for Last Observation Carried Forward

## Problems

Counterbalancing these advantages of LOCF are several disadvantages. First, because all the missing values for an individual are replaced with the same number, the within-subject variability is artificially reduced. In turn, this reduces the estimate of the error and, because the within-person error contributes to the denominator of any statistical test, it increases the likelihood of finding significance. Thus, rather than being conservative, LOCF actually might have a liberal bias and might lead to erroneously significant results.

Second, just as LOCF assumes no additional improvement for patients in the treatment condition, it also assumes that those in the comparison group will not change after they drop out of the trial. However, for many conditions, there is a very powerful placebo effect. In trials involving patients suffering from depression, up to 40% of those in the placebo arm of the study show significant improvement; and in studies of pain, this effect can be even stronger. Consequently, in underestimating the amount of change in the control group, LOCF again might have a positive bias, favoring rejection of the null hypothesis.

Finally, LOCF should never be used when the purpose of the intervention is to slow the rate of decline. For example, the so-called memory-enhancing drugs slow the rate of memory loss for patients suffering from mild or moderate dementia. In Figure 1, the left axis would now represent memory functioning, and thus lower scores are worse. If a person drops out of the study, then LOCF assumes no additional loss of functioning, which biases the results in favor of the treatment. In fact, the more people who drop out of the study, and the earlier the drop-outs occur, the better the drug looks. Consequently, LOCF introduces a very strong liberal bias, which significantly overestimates the effectiveness of the drug.

## Alternatives

Fortunately, there are alternatives to LOCF. The most powerful is called growth curve analysis (also known as latent growth modeling, latent curve analysis, mixed-effects regression, hierarchical linear regression, and about half a dozen other names), which can be used for all people who have

at least three data points. In essence, a regression line is fitted to each person's data, and the slope and intercept of the line become the predictor variables in another regression. This allows one to determine whether the average slope of the line differs between groups. This does not preserve as many cases as LOCF, because those who drop out with fewer than three data points cannot be analyzed, but latent growth modeling does not introduce the same biases as does LOCF.

*David L. Streiner*

*See also* Bias; Latent Growth Modeling; Missing Data, Imputation of

**Further Readings**

Molnar, F. J., Hutton, B., & Fergusson, D. (2008). Does analysis using "last observation carried forward" introduce bias in dementia research? *Canadian Medical Association Journal, 179,* 751–753.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* New York: Oxford University Press.

# Latent Growth Modeling

Latent growth modeling refers to a set of procedures for conducting longitudinal analysis. Statisticians refer to these procedures as mixed models. Many social scientists label these methods as multilevel analyses, and the label of hierarchical linear models is used in education and related disciplines. These procedures can be useful with static data where an individual response might be nested in a family. Thus, a response might be explained by individual characteristics, such as personality traits, or by a family-level characteristic, such as family income.

Longitudinal applications differ from static applications in that there are repeated measurements of a variable for each individual. The repeated measurements are nested in the individual. Just as individuals in a family tend to be similar, repeated measurements for the same individual tend to be similar. This lack of independence is handled by mixed models.

This entry begins with discussions of fixed and random effects and of time-varying and time-invariant predictors. Next, approaches are described and an example of the modeling process is provided. Last, additional extensions of latent growth modeling and its use in future research are examined.

## Fixed Versus Random Effects

To understand growth modeling, one needs to understand the concepts of fixed effects and random effects. In ordinary least-squares regression, a fixed intercept and a slope for each predictor are estimated. In growth modeling, it is often the case that each person has a different intercept and slope, which are called random effects. Consider a growth model of marital conflict reported by the mother across the first 12 months after the birth of a couple's first child. Conflict might be measured on a 0 to 10 scale right after the birth and then every 2 months for the first year. There are 7 time points (0, 2, . . . , 12) and a regression of the conflict scores on time might be done. Hypothetical results appear in Figure 1 in the graph labeled *Both Intercept and Slope Are Fixed*. The results reflect an intercept $\beta_0 = 2.5$ and a slope $\beta_1 = 0.2$. Thus, the conflict starts with an initial level of 2.5 and increases by 0.2 every 2 months. By the 12th month, the conflict would be moderate, $2.5 + 0.2 \times 12 = 4.9$. These results are fixed effects. However, women might vary in both their intercept and their slope.

In contrast, the graph in Figure 1 labeled *Random Intercept and Fixed Slope* allows for differences in the initial level and results in parallel lines. Mother A has the same intercept and slope as the fixed-effects model, 2.5 and 0.2, respectively. All three have a slope of 0.2, but they vary in their intercept (starting point). This random intercept model, by providing for individual differences in the intercept, should fit the data for all the mothers better than the fixed model, but the requirement that all lines are parallel might be unreasonable. An alternative approach is illustrated in the graph labeled *Fixed Intercept and Random Slope*. Here, all the mothers have a fixed initial level of conflict, but they are allowed to have different slopes (growth rates).

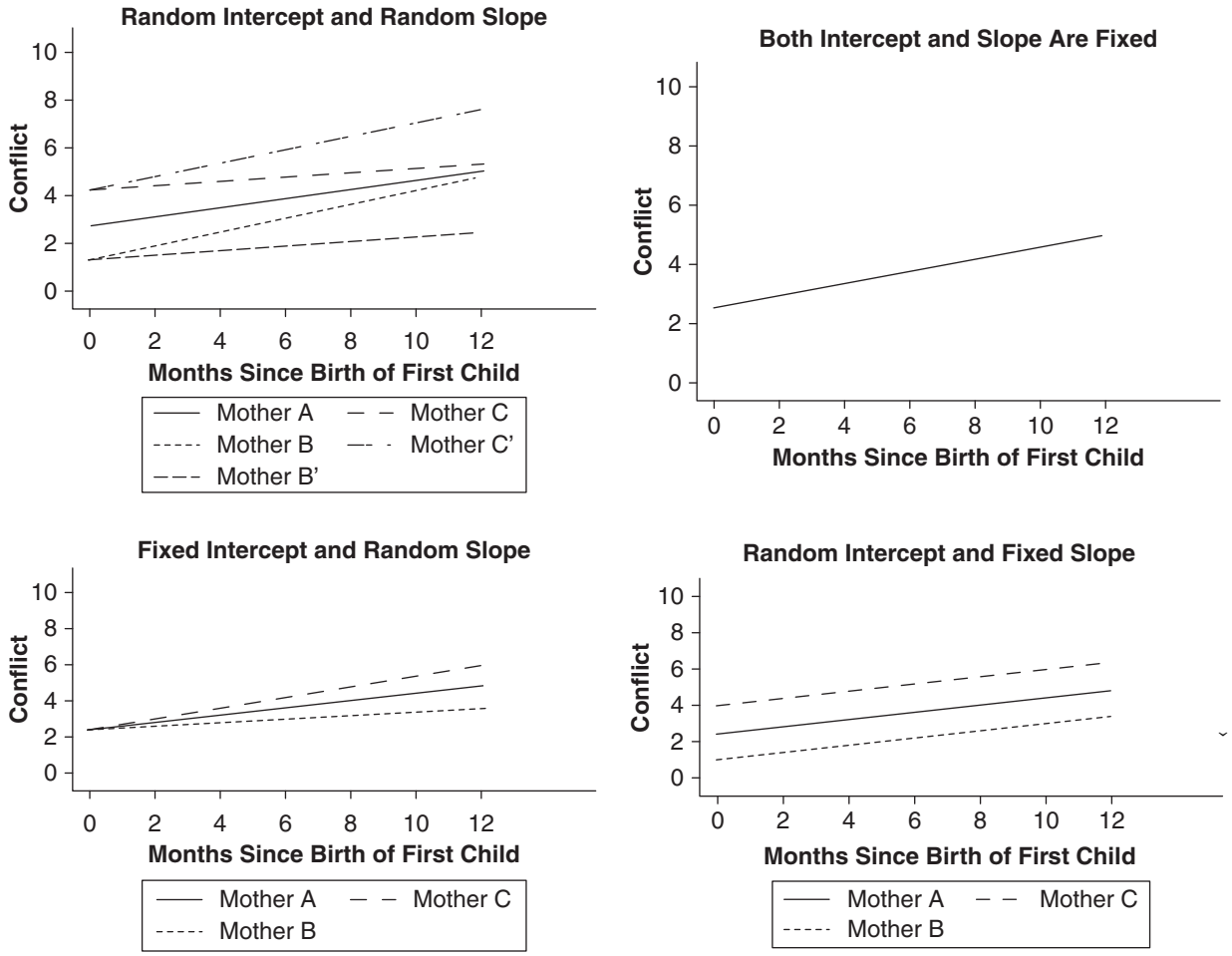



**Figure 1** Hypothetical Growth Curves for Marital Conflict in First 12 Months After Birth of First Child

Finally, the graph labeled *Random Intercept and Random Slope* allows each mother to have her own intercept and her own slope. This graph is more complicated than the fully fixed graph, but it might be a better fit to the data and seems realistic. This model with a random intercept and a random slope allows each mother to have a different initial level and a different growth trajectory.

## Time-Varying and Time-Invariant Predictors

When either or both the intercept and slope are treated as random effects, they are outcome variables that call for an explanation. Why do some mothers start with a high or low level of conflict? Why do mothers vary in their trajectory? The intercept and slope become dependent variables. Traditional regression analysis just estimates an intercept and slope; here, these estimates, when they are random effects, are treated as outcomes. Therefore, *change* in a variable is explained rather than just a static level of a variable.

Latent growth modeling allows researchers to explain such differences using two types of predictors (independent variables). These are called time-invariant covariates and time-varying covariates. First, the intercept or slope might depend on time-invariant covariates. These covariates are a constant value for each individual. For example, the mother's age at the birth and whether she was married at the time of birth are time invariant.

Time-varying covariates, by contrast, can have different values from one time to the next. Suppose 6 months after the birth of a child, the father attends a childcare workshop and becomes a more engaged parent. This would not predict her

intercept—that was at month 0, but it might predict a change in conflict at the 6th month and possibly at subsequent months.

In evaluation research, an important time-varying covariate is the fidelity of implementation. A researcher might have a 4-year program designed to improve math achievement. First, the researcher would estimate the fixed-effects growth curve. Second, he or she would test whether the random effects were significant. Third, the researcher would test whether time-invariant covariates (intervention vs. control group) were significant. Fourth, he or she would test whether year-to-year variability in the program fidelity had a significant effect. The fidelity of implementation each year is a time-varying covariate. It might explain why some years, some students scored much higher or lower on math achievement than their overall growth trajectory.

## Approaches and Example

Once one understands the concept of growth curve modeling, a statistical software package is needed to estimate the models. Most statisticians use commands available in Stata, SAS, or R to estimate mixed models. Researchers in education and related disciplines often use a program called HLM. Here, an alternative approach that focuses on the latent variable modeling approach to these questions is presented. This approach is an extension of structural equation modeling (SEM) in which the intercept and slope are conceptualized as latent growth factors. SEM software (Mplus, LISREL, EQS, and to a limited extent AMOS) have varying strengths for this approach. The models estimated here use the Mplus program because it is arguably the most comprehensive and the most rapidly developing. The Mplus program is available at the Mplus website, where there is extensive documentation and online training. To introduce latent growth modeling, data from the National Longitudinal Survey of Youth, 1997 is used. The focus is on the body mass index (BMI) of a panel of youth who were 12 years old in 1997. They were followed until they were 18 in 2003. There are limitations to the BMI measure, especially during adolescence, but the purpose is only to use it as an illustration of growth modeling.

### *Steps*

The first step involves examining the data in terms of distributions and level of measurement. Mplus and some other SEM packages can estimate models with or without assuming normality and regardless of the level of measurement. Mplus can also adjust standard errors for clustering, which would happen if a few youth from each of many schools were sampled. Alternatively, the researcher might want to incorporate school as an explanatory variable, perhaps including school-level variables (e.g., time spent in physical education curriculum).

It is useful to observe actual scores for a sample of observations. Figure 2 presents a random sample of 10 youth. A BMI of around 20 would be a normal weight; a BMI of 25 would be overweight. A BMI of 30 or higher would be obese. What is revealed in Figure 2? First, a random intercept is probably needed given that the sample of 10 youth varies from around a BMI of 12 to an initial BMI of around 30. What about the slope? These 10 observations do not tell much, although one might observe a general positive slope and that there is substantial variance in the growth rate. One youth who started at a BMI of 30 increased to one of 50 (morbidly obese) by the age of 18; other youth have a slope of about zero (flat). Based on these 10 observations, one would expect that both the intercept and the slope have a significant random effect.

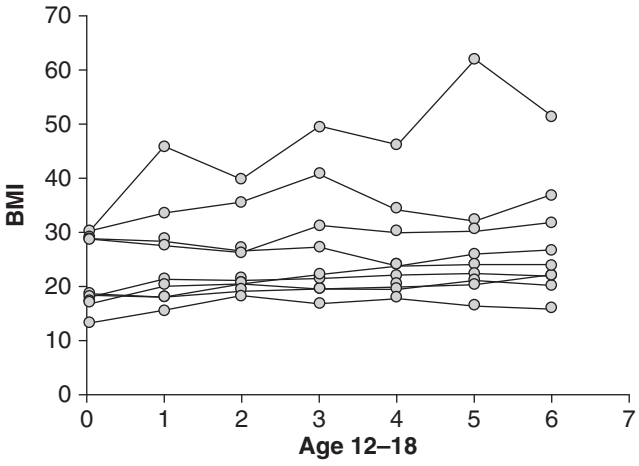



**Figure 2** Random Sample of the BMI Actual Scores for 10 Youth

The next step is developing a simple latent growth model. The simplest is a linear growth curve (called a curve, but a linear growth curve is actually a straight line requiring just an intercept factor and a linear slope factor). Figure 3 presents how a linear growth curve model can be drawn. This figure is simpler than it might appear. The oval labeled "intercept" is the *latent intercept growth factor.* It represents the initial level of the growth curve. Based on the sample of 10 youth, one might guess this will have a value of just more than 20. This value is the estimated initial $M_{\text{intercept}}$. The other oval, labeled "slope," is the *latent slope growth factor.* It represents how much the BMI increases (or decreases) each year. Using the sample of 10 youth, one might guess that this will be a small positive number, perhaps around 0.5. This value is the $M_{\text{slope}}$.

The sample of 10 youth indicates that there is variation around both the mean latent intercept growth factor and the mean latent slope growth factor. This variance is the random-effect component and is represented by the circles above the ovals, labeled $R_i$ (residual variance of latent intercept growth factor) and $R_s$ (residual variance of latent slope growth factor). If one of these variances is not significantly greater than zero, then that factor could be treated as a fixed effect. The curved line with an arrow at each end connecting $R_i$ and $R_s$ is the correlation of the latent intercept growth factor and latent slope growth factor. A positive correlation would indicate that people who start with a high BMI (intercept) have a more rapidly increasing BMI (slope) than people who start with a low BMI. Such a positive correlation is unfortunate both for youth who have a very high initial BMI (for whom a bigger slope is extremely problematic) and for youth with a very low BMI (who have a low or even negative slope).

How is the intercept identified? The intercept is often referred to as the constant. It represents the base value to which some amount is added or subtracted for each unit increase in the predictor. This constant base value is identified by having a fixed loading of 1.0 from it to each year's measurement of BMI. These lines are the dashed lines with an arrow going from the latent intercept to the individual BMI scores. The traditional meaning of a slope is how much a variable changes for each unit change in the independent variable. How is this translated to identify the latent slope growth factor? There are BMI measurements for 7 consecutive years, 1997–2003. Because each of these is a 1-year change, loadings of 0, 1, 2, 3, 4, 5, and 6 can be used, as illustrated by the solid lines with an arrow going from the latent slope growth factor to each year's measurement of BMI. Other fixed loadings might be appropriate. If no data were collected in 2000 and 2002, there would be five waves and loadings of 0, 1, 2, 4, and 6 could be used, simply dropping the missing years. One might want the intercept to represent the final level of the variable and use loadings of −6, −5, −4, −3, −2, −1, and 0, or put the intercept in the middle using loadings of −3, −2, −1, 0, 1, 2, and 3. In the Mplus program, it is also possible for each participant to have a different time span between measurements. John's BMI98 (BMI measurement in 1998) might be 14 months after his first measurement if there was some delay in data collection. His BMI99 might be only 10 months after his second wave.

The observed measurements appear in the rectangular boxes and are labeled BMI97 to BMI03. Figure 3 has circles at the bottom labeled E97 to E03. These represent measurement error. SEM software varies widely in how it programs a figure like this. The key part of the program in Mplus is

```
intercept slope | BMI97@0 BMI98@1
BMI99@2 BMI@3 BMI@4 BMI@;5 BMI@6.
```

The first name, `intercept`, which could be anything such as `i` or `alpha`, will always be the intercept. The second name, `slope`, which could be anything such as `s` or `beta`, will always be the linear slope. The logical "or bar |" tells the program this is a growth curve. Each path from the intercept to the observed scores must be 1.0. One only needs to specify the loadings for the slope (`BMI97` is set at 0, `BMI98` is set at 1, etc.). The Mplus program reads this single line and knows that the model in Figure 3 is being estimated. It is possible to override the assumptions of the program, such as specifying that the residual is not correlated or that some the measurement errors are correlated. These assumptions depend on research goals and hypotheses.

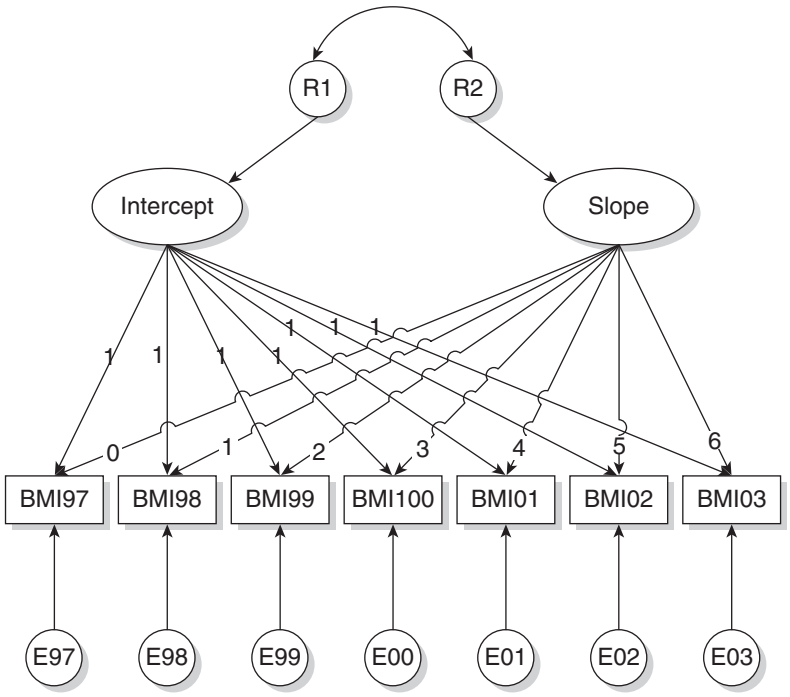



**Figure 3** A Linear Growth Curve Model of BMI Between Age 12 and 18

After running the program, information that helps in evaluating this simple growth curve is obtained. First, a chi-square of 268.04, $p < .000$, is obtained. This tests whether the simple growth curve fits the data perfectly. The significant chi-square says it does not. Because this is only a model, it is not surprising that it does not fit the data perfectly. A model can be very good and useful without perfectly reproducing all the observed data.

How close of a fit does the model in this example have? Among the many possible measures of closeness of fit, there is a comparative fit index (CFI) of .98, a root mean square error of approximation (RMSEA) of .078, and a standardized mean square residual (SRMR) of .051. The CFI should be more than .95, the RMSEA should be less than .06, and the SRMR should be less than .08. The results in this example are not too bad, but the RMSEA is probably the most respected measure of a close fit and it is too large.

What do the results reveal about the growth trajectory? The mean intercept is 21.035. The mean slope is .701. These are unstandardized values, and one needs to be careful how one interprets them. The estimated BMI $= 21.035 + .701 \times$ Year. Thus, one can estimate a youth has a BMI of 21.035 ($z = 210.352$, $p < .001$) at the age of 12 (well within the normal range), but this grows by .701 ($z = 40.663$, $p < .001$) each year. By 2003, the estimated BMI is $21.035 + .701 \times 6 = 25.241$. Thus, by the time they are 18, the expected BMI is into the overweight range. An increase in BMI of .701 each year is substantial.

The variance of the intercept and slope represents the size of the random effects. The variance of the intercept is 15.051, $z = 25.209$, $p < .001$, and the variance of the slope is .255, $z = 14.228$, $p < .001$. Both of these values are statistically significant. The $z$ tests are problematic tests of significance because the variance can never be negative. A better test, not reported here, involves estimating two models: one model that fixes the variances at zero and a second model that allows the variances to be freely estimated. These models are then compared based on the difference in chi-square and degrees of freedom. If the model with the variances estimated has a significantly smaller chi-square (calculated by subtraction), then a random-effects model is appropriate. A more elaborate test would compare the model with fixed intercept variance first with a model with a random intercept, then a model with both a random intercept and a random slope. These model comparisons are recommended whenever the $z$ tests are only marginally significant.

Can the variances be interpreted? The variances can be converted to standard deviations by taking their square roots. The square root of $15.051 = 3.880$. This means two thirds of the youth have an intercept within one standard deviation of 21.035, and about 95% have an intercept within two standard deviations of 21.035. The same thing can be done for the variance of the slope. Its square root is 0.505. This shows that people can have a slope of $.701 \pm 1.10$, and this covers everything from a dramatic increase in BMI to an estimated

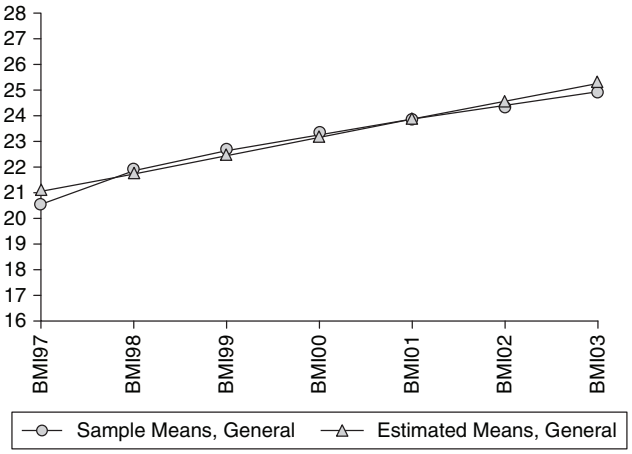



**Figure 4** Plot of BMI Growth Trajectory for Actual and Estimated Means

decrease in BMI. Clearly, there is an important random effect for both the intercept and the slope.

The covariance between the intercept and slope is .408, $z = 5.559$, $p < .001$ and the correlation is $r = .208$, $z = 5.195$, $p < .001$. [Mplus has a different test of significance for unstandardized and standardized coefficients, see Muthen (2007).] This correlation between the intercept and slope indicates exactly the area of concern, namely, that the youth who had the highest initial BMI scores also had the highest rate of growth in their BMI.

A plot of the linear latent growth curve can also be examined. In Figure 4, it can be observed that a straight line slightly overestimates the initial mean and slightly underestimates the final mean. This suggests that a quadratic can be added to the growth curve to capture the curvature in the observed data. When this is done, an excellent fit to the data is obtained.

### *How Many Waves of Data Are Needed?*

A latent growth model tries to reproduce the summary statistics describing the data. If there are just three waves of data, then the researcher would have three means, $M_{97}$, $M_{98}$, $M_{99}$; three variances, Var($BMI_{97}$), Var($BMI_{98}$), and Var($BMI_{99}$); and three covariances Cov($BMI_{97}$, $BMI_{98}$), Cov($BMI_{97}$, $BMI_{98}$), and Cov($BMI_{98}$, $BMI_{99}$) for a total of nine pieces of information he or she is trying to reproduce. How many parameters are being estimated? Consider how Figure 3 would look with three waves of data. The loadings for the intercept and the slope are fixed. Three error terms, E97, E98, E99, the $M_{\text{intercept}}$, the $M_{\text{slope}}$, variances, $R_i$ and $R_s$, as well as their covariance would need to be estimated. This gives a total of eight parameters to be estimated. Degrees of freedom are calculated by subtracting the number of parameters being estimated, eight, from the number of pieces of information obtained, nine, meaning there is just one degree of freedom. This does not provide a very rigorous test of a model, but it demonstrates that it is possible to estimate a linear growth curve with just three waves of data.

What if there were four waves? If one counted the means, variances, and covariances, there would be 14 pieces of information instead of 9. However, only one more parameter, E00, would be estimated, so there would be $14 - 9 = 5$ degrees of freedom. This gives a better test of a linear model. It also allows a quadratic term to be estimated to fit a curve. Adding a quadratic adds four parameters: $M_{\text{quadratic}}$, $R_Q$, and the covariances of the quadratic with both the intercept and the linear growth factors. It is good to have four waves of data for a linear growth curve, although three is the minimum and it is good to have at least five waves of data for a nonlinear growth curve, although four is the minimum.

### *Time-Invariant Covariates*

Whenever there is a significant variance in the intercept or slope, these random effects should be explained. For example, whites and nonwhites might be compared on their BMI. In this example, after dropping Asians and Pacific Islanders, the nonwhites are primarily African Americans and Latinos. Whites might have a lower intercept and a flatter slope then nonwhites in their BMI. If this were true, then race/ethnicity would explain a portion of the random effects. Race/ethnicity is a time-invariant covariate.

Consider emotional problems as a covariate that might explain some of the variance in the random effects. If this is measured at age 12 and not again, it would be treated as a time-invariant covariate. Children who have a high score on emotional problems at age 12 might have a different growth trajectory than children who have a low

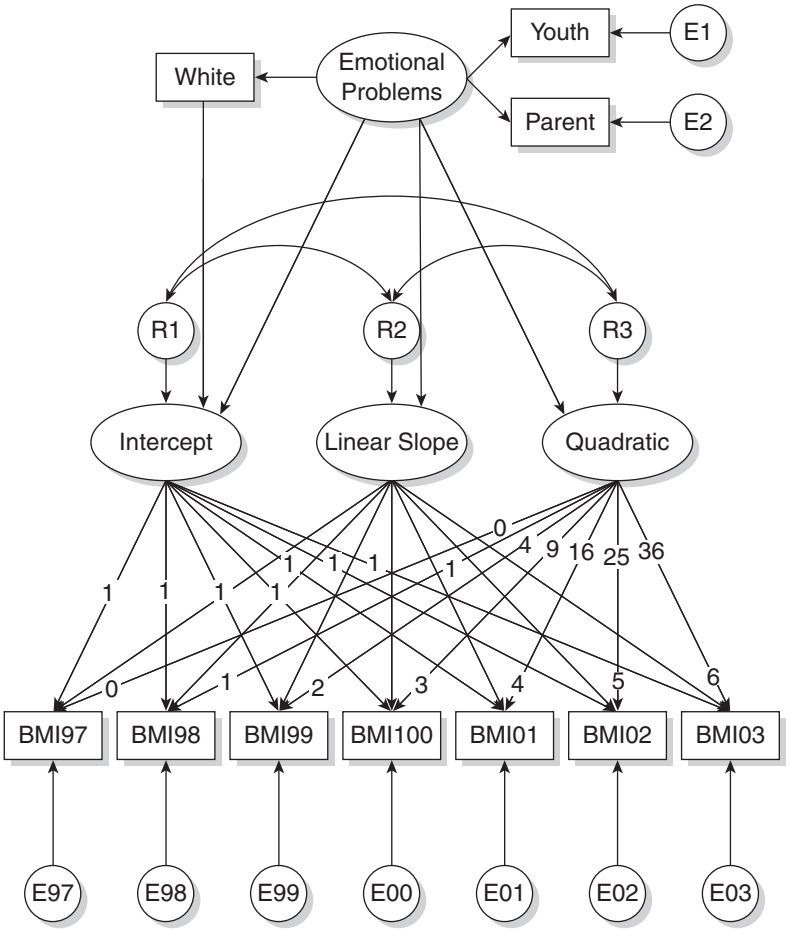



**Figure 5** Latent Growth Curve With Two Time-Invariant Covariates

score on emotional problems. Alternatively, if emotional problems are measured at each wave, it would be a time-varying covariate. In this section, race/ethnicity and age 12 emotional problems as time-invariant covariates are considered. There are other time-invariant covariates that should be considered, all of which would be measured just one time when the youth was 12 years old. For example, mothers' and fathers' BMI, knowledge of food choices, proximity of home to a fast-food restaurant, and many other time-invariant covariates could be measured.

Figure 5 shows the example model with these two covariates added. This has been called the conditional latent trajectory modeling because the initial level and trajectory (slope) are conditional on other variables. White is a binary variable coded 0 for nonwhite and 1 for white. It is in a rectangle like the measures of BMI because it is a single score. The emotional problems variable is presented in an oval because it is a latent variable (factor) that has two indicators, the youths' report and the parents' report. Both emotional problems and white race have lines with arrows going from them to the intercept and slope growth factors. This figure also has a quadratic slope factor with the loadings on BMI97 to BMI03 coded with the square of the corresponding loadings for the linear slope factor.

The results of this model can be used to test whether the trajectory is conditional on these two time invariant covariates. One also could use the $R^2$ for the intercept and slope factors to see how much variance in the random effects these covariate can explain.

### *Time-Varying Covariates*

In Figure 6, time-invariant covariates are represented by the rectangle labeled W. This represents a vector of possible time-invariant covariates that will influence the growth trajectory. It is possible to extend this to include time-varying covariates. Time-varying covariates either are measured after the process has started or have a value that changes (hours of nutrition education or level of program fidelity) from wave to wave. Although output is not shown, Figure 5 illustrates the use of time-varying covariates. In Figure 6, the time-varying covariates A1 to A6 might be the number of hours of curriculum devoted to nutrition education each year.

Time-varying covariates do not have a direct influence on the intercept or slope. For example, the amount of nutrition education youth received in 2003 could influence neither their initial BMI in 1997 nor their growth rate in earlier years. Instead, the hours of curriculum devoted to nutrition education each year would provide a direct effect on their BMI that year. A year with a strong

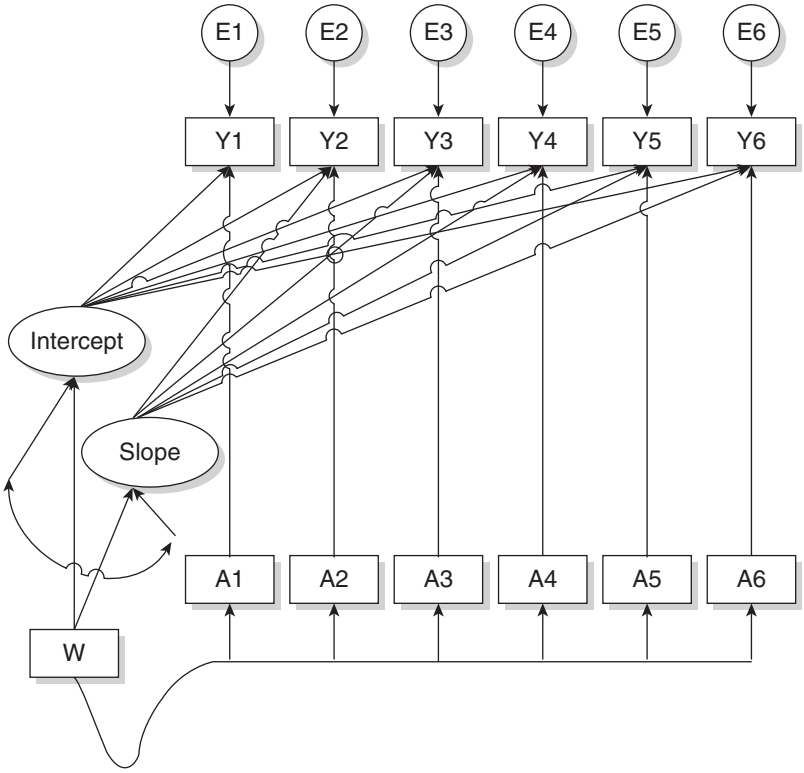



**Figure 6** Latent Growth Model With Time-Invariant and Time-Variant Covariates

curriculum might result in a lower BMI, and a year with little or no curriculum might result in a higher BMI. Thus, yearly curriculum would be explaining departures from the overall growth trajectory rather than the trajectory itself. To explain the random effects even more, additional paths could be added such as from the curriculum scores to the BMI in subsequent years.

## Extensions

Latent growth modeling is a rapidly developing area, and researchers have only scratched the surface of the potential it has for social science research. Along with the programs and data for the models discussed so far, several extensions are available, a few of which are mentioned here.

The latent growth curve model could be a part of a larger model. For example, the initial level and rate of change in BMI could lead to distal outcomes. Youth who have high initial BMIs and for whom this value is rapidly increasing might have long-term problems in careers, family relations, their role as parents, and health risks. One could add these distal variables with direct effects coming from the intercept and slope growth factors to these distal variables.

Parallel processes are possible. For example, one might have a growth curve for growth in conflict between parents and a parallel growth curve for growth in conflict the youth has with peers or teachers. High initial parental conflict could lead to a steeper slope for conflict the youth has with peers or teachers. The level of initial conflict and the growth in conflict for both variables could have long-term effects on a distal outcome, such as the youths' marital or parenting relations after they get married and have children.

Also, growth curves need not be limited to continuous variables. Growth could be a binary variable using a logit model where the growth in the probability of some outcome is estimated. For example, a school might introduce a program designed to stop smoking. Initially, there might be a high probability of smoking, but this probability might decrease across the course of the program. Time-variant (e.g., fidelity) and time-invariant covariates (e.g., gender) could be included in the model of this process to identify what predicts the program's effectiveness.

Other times, a count variable might be predicted. Whether a smoking cessation intervention eliminates smoking, it might reduce the number of cigarettes students smoke. Here, a Poisson or negative binomial model might be used to fit the data. Other times, researchers might be interested in both the binary growth curve and the count growth curve. These could be treated as parallel growth curves. It might be that some time-invariant and some time-variant covariates predict the binary or the count components differentially. Peers who smoke might make it difficult to stop smoking (binary component), but peer influence might not

have as much effect on the number of cigarettes smoked (count component). Finding such distinctions can provide a much better understanding of an intervention's effectiveness and can give ideas for how to improve the intervention.

The last of many possibilities to be mentioned here involves applications of growth mixture models. A sample can be treated as representing a single population, when there might be multiple populations represented and these multiple populations might have sharp differences. If one were to create a growth curve of abusive drinking behavior from age 18 to 37, one will find a growth curve that generally increases from age 18 to 23 and then decreases after that. However, this overall growth model might not fit the population very well. Why? What about people who never drink? These people have an intercept at or close to zero and a flat slope that is near zero. This is an identifiable population for which the overall growth curve is inapplicable. What about alcoholics? They might be similar to the overall pattern up to age 23, but then they do not decrease. Mixture models seek to find clusters of people who have homogeneous growth trajectories. Bengt Muthen and Linda Muthen applied a growth mixture model and were able to identify three clusters of people that we can label the normative group, the nondrinkers, and the likely alcoholics. Once identifying group membership, a profile analysis can be performed to evaluate how these groups differ on other variables. The same intervention would not work for the alcoholic that works for the normative group, and it is not cost effective to have an intervention on the group of nondrinkers.

### Future Directions

Latent growth curve modeling is one of the most important advances in the treasure chest of research methods that has developed in the last 20 years. It allows researchers to focus on change, what explains the rate of change, and the consequences of change. It is applicable across a wide range of subject areas and can be applied to data of all levels of measurement. It is also an area of rapid development and will likely continue to change the way researchers work with longitudinal data.

*Alan C. Acock*

***See also*** Growth Curve; Structural Equation Modeling

### Further Readings

Barrett, P. (2006). Structural equation modeling: Adjudging model fit. *Personality and Individual Differences, 42,* 815–824.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. Hoboken, NJ: John Wiley.

Center for Human Resources. (2007). *NLSY97 user's guide: A guide to rounds 1–9 data*. Washington, DC: U.S. Department of Labor.

Curran, F. J., & Hussong, A. M. (2003). The use of latent trajectory models in psychopathology research. *Journal of Abnormal Psychology, 112,* 526–544.

Duncan, T. E., Duncan, S. C., & Strycker, L. A. (2006). *An introduction to latent variable growth curve modeling: Concepts, issues, and applications* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

*Latent growth curves*. Retrieved July 2, 2009, from http://oregonstate.edu/~acock/growth

*Mplus homepage*. Retrieved January 27, 2010, from http://www.statmodel.com

Muthen, B. (2007). *Mplus technical appendices: Standardized coefficients and their standard errors*. Los Angeles. Retrieved January 27, 2010, from http://statmodel.com/download/techappen.pdf

Muthen, B., & Muthen, L. (2000). The development of heavy drinking and alcohol-related problems from ages 18 to 37 in a U.S. national sample. *Journal of Studies of Alcohol, 70,* 290–300.

Schreiber, J., Stage, R., King, J., Nora, A., & Barlow, E. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *The Journal of Educational Research, 99,* 323–337.

Yu, C. Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes* (Unpublished dissertation). University of California, Los Angeles.

## Latent Variable

A latent variable is a variable that cannot be observed. The presence of latent variables, however, can be detected by their effects on variables that are observable. Most constructs in research are latent variables. Consider the psychological construct of anxiety, for example. Any single observable measure of anxiety, whether it is a self-report measure or an observational scale, cannot

provide a pure measure of anxiety. Observable variables are affected by measurement error. Measurement error refers to the fact that scores often will not be identical if the same measure is given on two occasions or if equivalent forms of the measure are given on a single occasion. In addition, most observable variables are affected by method variance, with the results obtained using a method such as self-report often differing from the results obtained using a different method such as an observational rating scale. Latent variable methodologies provide a means of extracting a relatively pure measure of a construct from observed variables, one that is uncontaminated by measurement error and method variance. The basic idea is to capture the common or shared variance among multiple observable variables or indicators of a construct. Because measurement error is by definition unique variance, it is not captured in the latent variable. Technically, this is true only when the observed indicators are (a) obtained in different measurement occasions, (b) have different content, and (c) have different raters if subjective scoring is involved. Otherwise, they will share a source of measurement error that can be captured by a latent variable. When the observed indicators represent multiple methods, the latent variables also can be measured relatively free of method variance. This entry discusses two types of methods for obtaining latent variables: exploratory and confirmatory. In addition, this entry explores the use of latent variables in future research.

## Exploratory Methods

Latent variables are linear composites of observed variables. They can be obtained by exploratory or confirmatory methods. Two common exploratory methods for obtaining latent variables are factor analysis and principal components analysis. Both approaches are exploratory in that no hypotheses typically are proposed in advance about the number of latent variables or which indicators will be associated with which latent variables. In fact, the full solutions of factor analyses and principal components analyses have as many latent variables as there are observed indicators and allow all indicators to be associated with all latent variables. What makes exploratory methods useful is when most of the shared variance among observed indicators can be accounted for by a relatively small number of latent variables.

The measure of the degree to which an indicator is associated with a latent variable is the indicator's loading on the latent variable. An inspection of the pattern of loadings and other statistics is used to identify latent variables and the observed variables that are associated with them. Principal components are latent variables that are obtained from an analysis of a typical correlation matrix with 1s on the diagonal. Because the variance on the diagonal of a correlation matrix is a composite of common variance and unique variance including measurement error, principal components differ from factors in that they capture unique as well as shared variance among the indicators. Because all variance is included in the analysis and exact scores are available, principal components analysis primarily is useful for "boiling down" a large number of observed variables into a manageable number of principal components.

In contrast, the factors that result from factor analysis are latent variables obtained from an analysis of a correlation matrix after replacing the 1s on the diagonal with estimates of each observed variable's shared variance with the other variables in the analysis. Consequently, factors capture only the common variance among observed variables and exclude measurement error. Because of this, principal factor analysis is better for exploring the underlying factor structure of a set of observed variables.

## Confirmatory Methods

Latent variables can also be identified using confirmatory methods such as confirmatory factor analysis and structure equation models with latent variables, and this is where the real power of latent variables is unleashed. Similar to exploratory factor analysis, confirmatory factor analysis captures the common variance among observed variables. However, predictions about the number of latent variables and about which observed indicators are associated with them are made a priori (i.e., prior to looking at the results) based on theory and prior research. Typically, observed indicators are only associated with a single latent variable when confirmatory methods are used. The a priori predictions about the number of latent variables and which indicators are associated with them can be

tested by examining how well the specified model fits the data. The use of a priori predictions and the ability to test how well the model fits the data are important advantages over exploratory methods for obtaining latent variables.

One issue that must be considered is that it is possible to propose a confirmatory latent variable model that does not have a unique solution. This is referred to as an underidentified model. Underidentified confirmatory factor analysis models can usually be avoided by having at least three observed indicators for a model with a single latent variable and at least two observed indicators for each latent variable in a model with two or more latent variables, provided that they are allowed to be correlated with one another.

### Future Research

Until recently, latent variables have been continuous rather than categorical variables. With the development of categorical latent variables, latent variables are proving to be a powerful new tool for identifying mixtures of different kinds of people or subtypes of various syndromes. Models with categorical latent variables (e.g., mixture models) are replacing cluster analysis as the method of choice for categorizing people. Other recent advances in latent variable analysis include latent growth models, transition mixture models, and multi-level forms of all of the models described previously.

*Richard Wagner, Patricia Thatcher Kantor, and Shayne Piasta*

***See also*** Confirmatory Factor Analysis; Exploratory Factor Analysis; Principal Components Analysis; Structural Equation Modeling

#### Further Readings

Hancock, G. R., & Mueller, R. O. (2006). *Structural equation modeling: A second course*. Greenwich, CT: Information Age Publishing.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford.

Schumacker, R. E., & Lomax, R. G. (1996). *A beginner's guide to structural equation modeling*. Mahwah, NJ: Lawrence Erlbaum.

## Latin Square Design

In general, a Latin square of order $n$ is an $n \times n$ square such that each row (and each column) is a permutation (or an arrangement) of the same $n$ distinct elements. Suppose you lead a team of four chess players to play four rounds of chess against another team of four players. If each player must play against a different player in each round, a possible schedule could be as follows:

(a)

| Team A | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Round 1 | 1 | 3 | 4 | 2 |
| Round 2 | 4 | 2 | 1 | 3 |
| Round 3 | 2 | 4 | 3 | 1 |
| Round 4 | 3 | 1 | 2 | 4 |

Here, we assumed that the players are numbered 1 to 4 in each team. For instance, in round 1, player 1 of team A will play 1 of team B, player 2 of team A will play 3 of team B, and so on.

Suppose you like to test four types of fertilizers on tomatoes planned in your garden. To reduce the effect of soils and watering in your experiment, you might choose 16 tomatoes planned in a $4 \times 4$ square (i.e., four rows and four columns), such that each row and each column has exactly one plan that uses one type of the fertilizers. Let the fertilizers denoted by A, B, C, and D, then one possible experiment is denoted by the following square on the left.

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | A | C | D | B |
| 2 | D | B | A | C |
| 3 | B | D | C | A |
| 4 | C | A | B | D |

**(a)**

| ⋆ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | 1 | 3 | 4 | 2 |
| 2 | 4 | 2 | 3 | 1 |
| 3 | 2 | 4 | 3 | 1 |
| 4 | 3 | 1 | 2 | 4 |

**(b)**

This table tells us that the plant at row 1 and column 1 will use fertilizer A, the plant at row 1 and column 2 will use fertilizer C, and so on. If we rename A to 1, B to 2, C to 3, and D to 4, then we obtain a square in the (b) portion of the table, which is identical to the square in the chess schedule.

In mathematics, all three squares in the previous section (without the row and column names) are called a *Latin square* of order four. The name Latin square originates from mathematicians of the 19th century like Leonhard Euler, who used Latin characters as symbols.

## Various Definitions

One convenience of using the same set of elements for both the row and column indices and the elements inside the square is that we might regard $\star$ as a function $\star$ defined on the set {1,2,3,4}: $\star(1,1) = 1$, $\star(1,2) = 3$, and so on. Of course, we might write it as $1 \star 1 = 1$ and $1 \star 2 = 3$, and we can also show that $(1 \star 3) \star 4 = 4$. This square provides a definition of $\star$, and the square (b) in the previous section is called "the multiplication table" of $\star$.

In mathematics, if the multiplication table of a binary function, say $*$, is a Latin square, then that function, together with the set of the elements, is called *quasigroup*. In contrast, if $*$ is a binary function over the set

$$S = \{1, 2, \ldots, n\}$$

and satisfies the following properties, then the multiplication table $*$ is a Latin square of order $n$.

For all elements $w, x, y, z \in S$,

$$(x * w = z, x * y = z) \Rightarrow w = y: \quad \text{the left-cancellation law;} \tag{1}$$

$$(x * y = z, x * y = z) \Rightarrow w = x: \quad \text{the right-cancellation law;} \tag{2}$$

$$(x * y = w, x * y = z) \Rightarrow w = z: \quad \text{the unique-image property.} \tag{3}$$

The left-cancellation law states that no symbol appears in any column more than once, and the right-cancellation law states that no symbol appears in any row more than once. The unique-image property states that each cell of the square can hold at most one symbol.

A Latin square can be also defined by a set of triples. Let us look at the first Latin square [i.e. square (a)].

| | | | | | |
|---|---|---|---|---|---|
| 1 | 2 | 3 | 1 | 2 | 3 |
| 2 | 3 | 1 | 3 | 1 | 2 |
| 3 | 1 | 2 | 2 | 3 | 1 |
| | **(a)** | | | **(b)** | |
| 1 | 3 | 2 | 1 | 3 | 2 |
| 2 | 1 | 3 | 3 | 2 | 1 |
| 3 | 2 | 1 | 2 | 1 | 3 |
| | **(c)** | | | **(d)** | |

The triple representation of square (a) is

$$\{\langle 1, 1, 1\rangle, \langle 1, 2, 2\rangle, \langle 1, 3, 3\rangle, \langle 2, 1, 2\rangle, \langle 2, 2, 3\rangle, \langle 2, 3, 1\rangle, \langle 3, 1, 3\rangle, \langle 3, 2, 1\rangle, \langle 3, 3, 2\rangle\}.$$

The meaning of a triple $\langle x, y, z\rangle$ is that the entry value at row $x$ and column $y$ is $z$ (i.e., $x * y = z$, if the Latin square is the multiplication table of $*$). The definition of a Latin square of order $n$ can be a set of $n^2$ integer triples $\langle x, y, z\rangle$, where $1 \leq x, y, z \leq n$, such that

- All the pairs $\langle x, z\rangle$ are different (the left-cancellation law).
- All the pairs $\langle y, z\rangle$ are different (the right-cancellation law).
- All the pairs $\langle x, y\rangle$ are different (the unique-image property).

The quasigroup definition and the triple representation show that rows, columns, and entries play similar roles in a Latin square. That is, a Latin square can be viewed as a cube, where the dimensions are row, column, and entry value. If two dimensions in the cube are fixed, then there will be

exactly one number in the remaining dimension. Moreover, if a question or a solution exists for a certain kind of Latin squares for one dimension, then more questions or solutions might be generated by switching dimensions.

## Counting Latin Squares

Let us first count Latin squares for some very small orders. For order one, we have only one Latin square. For order two, we have two (one is identical to the other by switching rows or columns). For order three, we have 12 (four of them are listed in the previous section). For order four, we have 576 Latin squares. A Latin square is said be *reduced* if in the first row and the first column the elements occur in natural order. If $R(n)$ denotes the number of reduced Latin square of order $n$, then the total number of Latin squares, $N(n)$, is given by the following formula:

$$N(n) = n!(n-1)!R(n),$$

where $n! = n*(n-1)*(n-2)\ldots*2*1$ is the factorial of $n$. From this formula, we can deduce that the numbers of reduced Latin squares are 1,1,3,4, for order 1,2,3,4, respectively.

Given a Latin square, we might switch rows to obtain another Latin square. For instance, Latin square (b) is obtained from (a) by switching rows 2 and 3. Similarly, (c) is obtained from (a) by switching columns 2 and 3, and (d) is obtained from (a) by switching symbols 2 and 3. Of course, more than two rows might be involved in a switch. For example, row 1 can be changed to row 2, row 2 to row 3, and row 3 to row 1. Two Lain squares are said be to *isotopic* if one becomes the other by switching rows, columns, or symbols. Isotopic Latin squares can be merged into one class, which is called the *isotopy class*.

The following table gives the numbers of nonisotopic and reduced Latin squares of order up to ten.

## Conjugates of a Latin Square

Given a Latin square, if we switch row 1 with column 1, row 2 with column 2, ..., and row $n$ with column $n$, then we obtain the *transpose* of the original Latin square, which is also a Latin square. If the triple representation of this Latin square is $S$, then the triple representation of the transpose is $T = \{\langle x_2,x_1,x_3\rangle | \langle x_1,x_2,x_3\rangle \in S\}$. For the three elements $\langle x_1,x_2,x_3\rangle$, we have six permutations, each of which will produce a Latin square, which is called a *conjugate*. Formally, the transpose $T$ is the (2,1,3) conjugate;

$U = \{\langle x_1, x_3, x_2\rangle | \langle x_1, x_2, x_3\rangle \in S\}$ is the $(1, 3, 2)$ conjugate of $S$;
$V = \{\langle x_2, x_1, x_3\rangle | \langle x_1, x_2, x_3\rangle \in S\}$ is the $(2, 1, 3)$ conjugate of $S$;
$W = \{\langle x_2, x_3, x_1\rangle | \langle x_1, x_2, x_3\rangle \in S\}$ is the $(3, 1, 2)$ conjugate of $S$;
$X = \{\langle x_3, x_1, x_3\rangle | \langle x_1, x_2, x_3\rangle \in S\}$ is the $(3, 2, 1)$ conjugate of $S$;
$Y = \{\langle x_3, x_2, x_1\rangle | \langle x_1, x_2, x_3\rangle \in S\}$ is the $(1, 2, 3)$ conjugate of $S$.

Needless to say, the (1,2,3) conjugate of $S$ is $S$ itself. The six conjugates of a small Latin square are as follows:

| | | | | | |
|---|---|---|---|---|---|
| 1 4 2 3 | 1 2 4 3 | 1 3 2 4 | 1 2 4 3 | 1 3 4 2 | 1 3 2 4 |
| 2 3 1 4 | 4 3 1 2 | 2 4 1 3 | 3 4 2 1 | 3 1 2 4 | 3 1 4 2 |
| 4 1 3 2 | 2 1 3 4 | 4 2 3 1 | 2 1 3 4 | 2 4 3 1 | 4 2 3 1 |
| 3 2 4 1 | 3 4 2 1 | 3 1 4 2 | 4 3 1 2 | 4 2 1 3 | 2 4 1 3 |
| (a) | (b) | (c) | (d) | (e) | (f) |

(a) a Latin square, (b) its (2,1,3) conjugate, (c) its (3,2,1) conjugate, (d) its (2,3,1) conjugate, (e) its (1,3,2) conjugate, and (f) its (3,1,2) conjugate.

Some conjugates might be identical to the original Latin square. For instance, a symmetric Latin square [like square (a) of order three] is identical to its (2,1,3) conjugate. Two Latin squares are said be *parastrophic* if one of them is a conjugate of the other. Two Latin squares are

| $n$ = | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| *Nonisotopic* | 1 | 1 | 1 | 2 | 2 | 22 | 563 | 1,676,267 | 115,618,721,533 | 208,904,371,354,363,006 |
| *Reduced* | 1 | 1 | 1 | 4 | 56 | 9,408 | 16,942,080 | 535,281,401,856 | $3.77 \times 10^{18}$ | $7.58 \times 10^{25}$ |

said be *paratopic* if one of them is isotopic to a conjugate of the other. Like the isotopic relation, both the parastrophic and paratopic relations are equivalence relations. For orders of no less than six, the number of nonparatopic Latin squares is less than that of nonisotopic Latin squares.

## Transversals of a Latin Square

Given a Latin square of order *n,* a *transversal* of the square is a set *S* of *n* entries, one selected from each row and each column such that no two entries contain the same symbol. For instance, four transversals of the Latin square (a) are shown in the following as (b) through (e):

| | | | |
|---|---|---|---|
| 1 | 3 | 4 | 2 |
| 4 | 2 | 1 | 3 |
| 2 | 4 | 3 | 1 |
| 3 | 1 | 2 | 4 |

(a)

| | | | |
|---|---|---|---|
| 1 | | | |
| | | | 3 |
| | 4 | | |
| | | 2 | |

(b)

| | | | |
|---|---|---|---|
| | 3 | | |
| | | 1 | |
| 2 | | | |
| | | | 4 |

(c)

| | | | |
|---|---|---|---|
| | | 4 | |
| | 2 | | |
| | | | 1 |
| 3 | | | |

(d)

| | | | |
|---|---|---|---|
| | | | 2 |
| 4 | | | |
| | | 3 | |
| | 1 | | |

(e)

If we overlap the previous four transversals, we obtain the original Latin square. This set of transversals is called a *transversal design* (of index one). The total number of transversals for the Latin square (a) is eight. The four other transversals include the two diagonals, the entry set {(1,2), (2,1), (3,4), (4,3)} and the entry set {(1,3), (2,4), (3,1), (4,2)}. This set of four transversals also consists of a transversal design. Because the set of all eight transversals can be partitioned into two traversal designs, this Latin square has a *resolvable* set of transversals.

The research problem concerning transversals includes finding the maximal numbers of transversals, resolvable transversals, or transversal designs for each order. The maximal number of transversals for orders under nine are known; for orders greater than nine, only lower and upper bounds are known.

## Partial Latin Squares

A *partial Latin square* is a square such that each entry in the square contains either a symbol or is empty, and no symbol occurs more than once in any row or column. Given a partial Latin square, we often ask whether the empty entries can be filled to form a complete Latin square. For instance, it is known that a partial Latin square of order $n$ with at most $n - 1$ filled entries can always be completed to a Latin square of order $n$. If in a partial Latin square, $1*1 = 1$ and $2*2 = 2$, then we cannot complete this Latin square if the order is just two.

The most interesting example of the Latin square completion problem is perhaps the Sudoku puzzle, which appears in numerous newspapers and magazines. The most popular form of Sudoku has a $9 \times 9$ grid made up of nine $3 \times 3$ subgrids called "regions." In addition to the constraints that every row and every column is a permutation of 1 through 9, each region is also a permutation of 1 through 9. The less the number of filled entries (also called *hints*) in a Sudoku puzzle, the more difficult the puzzle. Gordon Royle has a collection of 47,386 distinct Sudoku configurations with exact 17 filled cells. It is an open question whether 17 is the minimum number of entries for a Sudoku puzzle to have a unique solution. It is also an open question whether Royle's collection of 47,386 puzzles is complete.

## Holey Latin Squares

Among partial Latin squares, people are often interested in Latin squares with holes, that is, some subsquares of the square are missing. The existence of these holey Latin squares is very useful in the construction of Latin squares of large orders.

Suppose $H$ is a subset of $\{1, 2, \ldots, n\}$; a *holey Latin square* of order $n$ with hole $H$ is a set of $n^2 - |H|^2$ integer triples $\langle x, y, z\rangle$, $1 \leq x, y, z \leq n$, such that

1. All $\langle x,y\rangle$ are distinct and at most one of them is in $H$.
2. All $\langle x,z\rangle$ are distinct and at most one of them is in $H$.

3. All $\langle y,z \rangle$ are distinct and at most one of them is in $H$.

Let $H = \{1,2\}$, and holey Latin squares of orders 4 and 5 with hole $H$ are given in the following as squares (a) and (b).

| | | | |
|---|---|---|---|
| | | 4 | 3 |
| | | 3 | 4 |
| 4 | 3 | 1 | 2 |
| 3 | 4 | 2 | 1 |

(a)

| | | | | |
|---|---|---|---|---|
| | | 4 | 3 | 5 |
| | | 3 | 5 | 4 |
| 3 | 4 | 5 | 2 | 1 |
| 5 | 3 | 1 | 4 | 2 |
| 4 | 5 | 2 | 1 | 3 |

(b)

| | | | | | |
|---|---|---|---|---|---|
| | | 5 | 6 | 3 | 4 |
| | | 6 | 5 | 4 | 3 |
| 5 | 6 | | | 2 | 1 |
| 6 | 5 | | | 1 | 2 |
| 3 | 4 | 2 | 1 | | |
| 4 | 3 | 1 | 2 | | |

(c)

Similarly, we might have more than one hole. For instance, (c) is a holey Latin square of order 6, with holes {1,2}, {3,4}, and {5,6}. Holey Latin squares with multiple holes (not necessarily mutually disjoint or same size) can be defined similarly using the triple representation. Obviously, holey Latin squares are a special case of partial Latin squares. A necessary condition for the existence of a holey Latin square is that the hole size cannot exceed the half of the order. There are few results concerning the maximal number of holey Latin squares for various orders.

## Orthogonal Latin Squares

Given a Latin square of order $n$, there are $n^2$ entry positions. Given a set of $n$ symbols, there are $n^2$ distinct pairs of symbols. If we overlap two Latin squares of order $n$, we obtain a pair of symbols at each entry position. If the pair at each entry position is distinct compared to the other entry positions, we say the two Latin squares are *orthogonal*. The following are some pairs of orthogonal Latin squares of small orders. There is a pair of numbers in each entry; the first of these comes from the first square and the second from the other square.

| | | |
|---|---|---|
| 1, 1 | 2, 3 | 3, 2 |
| 2, 2 | 3, 1 | 1, 3 |
| 3, 3 | 1, 2 | 2, 1 |

(a)

| | | | |
|---|---|---|---|
| 1, 1 | 2, 3 | 3, 4 | 4, 2 |
| 2, 2 | 1, 4 | 4, 3 | 3, 1 |
| 3, 3 | 4, 1 | 1, 2 | 2, 4 |
| 4, 4 | 3, 2 | 2, 1 | 1, 3 |

(b)

| | | | | |
|---|---|---|---|---|
| 1, 1 | 4, 5 | 2, 4 | 5, 3 | 3, 2 |
| 5, 4 | 2, 2 | 1, 5 | 3, 1 | 4, 3 |
| 4, 2 | 5, 1 | 3, 3 | 2, 5 | 1, 4 |
| 3, 5 | 1, 3 | 5, 2 | 4, 4 | 2, 1 |
| 2, 3 | 3, 4 | 4, 1 | 1, 2 | 5, 5 |

(c)

The orthogonality of Latin squares is perhaps the most important property in the study of Latin squares. One problem of great interests is to prove the existence of a set of mutually orthogonal Latin squares (MOLS) of certain order.

This can be demonstrated by Euler's 36 Officers Problem, in which one attempts to arrange 36 officers of 6 different ranks and 6 different regiments into a square so each line contains 6 officers of different ranks and regiments.

If the ranks and regiments of these 36 officers arranged in a square are represented, respectively, by two Latin squares of order six, then Euler's 36 officers problem asks whether two orthogonal Latin squares of order 6 exist. Euler went on to conjecture that such an $n \times n$ array does not exist for $n = 6$, and one does not exist whenever $n \equiv 2 \pmod{}$. This was known as the *Euler conjecture* until its disproof in 1959. At first, Raj C. Bose and Sharadchandra S. Shrikhande found some counterexamples; the next year, Ernest Tilden Parker, Bose, and Shrikhande were able to construct a pair of orthogonal order 10 Latin squares, and they provided a construction for the remaining even values of $n$ that are not divisible by 4 (of course, excepting $n = 2$ and $n = 6$). Today's computer software can find a pair of such Latin squares in no time. However, it remains a great challenge to find a set of three mutually orthogonal Latin squares of order 10.

Let $M(n)$ be the maximum number of Latin squares in a set of MOLS of order $n$. The following results are known: If $n$ is a prime power, that is, $n = p^e$, where $p$ is a prime, then $M(n) = n - 1$. For small $n > 6$ and $n$ is not a prime power, we do not know the exact value of $M(n)$ except the lower bounds as given in the following table:

| $n$ | 6 | 10 | 12 | 14 | 15 | 18 | 20 | 21 | 22 | 24 |
|---|---|---|---|---|---|---|---|---|---|---|
| $M(n)$ | 1 | $\geq 2$ | $\geq 5$ | $\geq 3$ | $\geq 4$ | $\geq 3$ | $\geq 4$ | $\geq 5$ | $\geq 3$ | $\geq 5$ |

MOLS can be used to design experiments. Suppose a drug company has four types of headache drugs, four types of fever drugs, and four types of cough drugs. To design a new cold medicine, the company wants to test the combinations of these three kinds of drugs. In a test, three drugs (not the same type) will be used simultaneously. Can we design only 16 tests so that every pair of drugs (not the same type) will be tested? The answer is yes, as we have a set of three MOLS of order four. A pair of MOLS is equivalent to a transversal design of index one.

People are also interested in whether a Latin square is orthogonal to its conjugate and the existence of mutually orthogonal holey Latin squares. For instance, for the two orthogonal squares of order 5 in the beginning of this section [i.e., (c)], one is the other's (2,1,3) conjugate. It has been known that a Latin square exists that is orthogonal to its (2,1,3), (1,3,2), and (3,2,1) conjugates for all orders except 2, 3, and 6; a Latin square exists that is orthogonal to its (2,3,1) and (3,1,2) conjugates for all orders except 2, 3, 4, 6, and 10. For holey Latin squares, the result is less conclusive.

### Applications

The two examples in the beginning of this entry show that Latin squares can be used for tournament scheduling and experiment design. This strategy has also been used for designing puzzles and tests. As a matching procedure, Latin squares relate to problems in graph theory, job assignment (or Marriage Problem), and, more recently, processor scheduling for massively parallel computer systems. Algorithms for solving the Marriage Problem are also used in linear algebra to reduce matrices to block diagonal form.

Latin squares have rich connections with many fields of design theory. A Latin square is also equivalent to a $(3,n)$ net, an orthogonal array of strength two and index one, a 1-factorization of the complete bipartite graph $K_{n,n}$, an edge-partition of the complete tripartite graph $K_{n,n,n}$ into triangles, a set of $n^2$ mutually nonattacking rooks on a $n \times n \times n$ board, and a single error-detecting code of word length 3, with $n^2$ words from an $n$-symbol alphabet.

*Hantao Zhang*

*Note*: Partially supported by the National Science Foundation under Grants CCR-0604205.

### Further Readings

Bennett, F., & Zhu, L. (1992). Conjugate-orthogonal Latin squares and related structures. In J. Dinitz & D. Stinson (Eds.), *Contemporary design theory: A collection of surveys*. New York: John Wiley.

Bose, R. C., & Shrikhande, S. S. (1960). On the construction of sets of mutually orthogonal Latin squares and falsity of a conjecture of Euler. *Transactions of the American Mathematical Society, 95,* 191–209.

Carley, A. (1890). On Latin squares. *Oxford Cambridge Dublin Messenger Mathematics, 19,* 135–137.

Colbourn, C. J. (1984). The complexity of completing partial Latin squares. *Discrete Applied Mathematics, 8,* 25–30.

Colbourn, C. J., & Dinitz, J. H. (Eds.). (1996). *The CRC handbook of combinatorial design*s. Boca Raton, FL: CRC Press.

Euler, L. (1849). Recherches sur une espèe de carr magiques. *Commentationes Arithmeticae Collectae, 2,* 302–361.

Evans, T. (1975). Algebraic structures associated with Latin squares and orthogonal arrays. *Proceedings of Conf. on Algebraic Aspects of Combinatorics, Congressus Numerantium, 13,* 31–52.

Hedayat, A. S., Sloane, N. J. A., & Stufken, J. (1999). *Orthogonal arrays: Theory and applications*. New York: Springer-Verlag.

Mandl, R. (1985). Orthogonal Latin squares: An application of experiment design to compiler testing. *Communications of the ACM, 28,* 1054–1058.

McKey, B. D., & McLeod, J. C. (2006). The number of transversals in a Latin square. *Des Codes Cypt, 40,* 269–284.

Royle, G. (n.d.). *Minimum Sudoku*. Retrieved January 27, 2010, from http://people.csse.uwa.edu.au/gordon/sudokumin.php

Tarry, G. (1900). Le probleme de 36 officiers. *Compte Rendu de l'Association Francaise Avancement Sciences Naturelle, 1,* 122–123.

Zhang, H. (1997). Specifying Latin squares in propositional logic. In R. Veroff (Ed.), *Automated reasoning and its applications: Essays in honor of Larry Wos*. Cambridge: MIT Press.

## Law of Large Numbers

The Law of Large Numbers states that larger samples provide better estimates of a population's

parameters than do smaller samples. As the size of a sample increases, the sample statistics approach the value of the population parameters. In its simplest form, the Law of Large Numbers is sometimes stated as the idea that bigger samples are better. After a brief discussion of the history of the Law of Large Numbers, the entry discusses related concepts and provides a demonstration and the mathematical formula.

## History

Jakob Bernoulli first proposed the Law of Large Numbers in 1713 as his "Golden Theorem." Since that time, numerous other mathematicians (including Siméon-Denis Poisson who first coined the term *Law of Large Numbers* in 1837) have proven the theorem and considered its application in games of chance, sampling, and statistical tests. Understanding the Law of Large Numbers is fundamental to understanding the essence of inferential statistics, that is, why one can use samples to estimate population parameters. Despite its primary importance, it is often not fully understood. Consequently, the understanding of the concept has been the topic of numerous studies in mathematics education and cognitive psychology.

## Sampling Distributions

Understanding the Law of Large Numbers requires understanding how sampling distributions differ for samples of various sizes. For example, if random samples of 10 men are drawn and their mean heights are calculated so that a frequency distribution of the mean heights can be created, a large amount of variability might be expected between those means. With the mean height of adult men in the United States at about 70 inches (5'10" or about 177 cm), some samples of 10 men could have means as high as 80 inches (6'6"), whereas others might be as low as 60 inches (5'0"). Although as the central limit theorem suggests the mean of the sampling distribution of the means will be equal to the population mean of 70 inches, the individual sample means will vary substantially. In samples of only 10 randomly selected men, it is easily possible to get an unusually tall group of 10 men or an unusually short group of men. Additionally, in such a small group, one outlier, for example who is 85 inches, can have a large effect on the sample mean. However, if samples of 100 men were drawn from the population, the means of those samples would vary less than the means from the samples of 10 men. It is much more difficult to select 100 tall men randomly from the population than it is to select 10 tall men randomly. Furthermore, if samples of 1,000 men are drawn, it is extremely unlikely that 1,000 tall men will be randomly selected. The mean heights for those samples would vary even less than the means from the samples of 100 men. Thus, as sample sizes increase, the variability between sample statistics decreases. The sample statistics from larger samples are, therefore, better estimates of the true population parameters.

## Demonstration

If a fair coin is flipped a million times, we expect that 50% of the flips will result in heads and 50% in tails. Imagine having five people flip a coin 10 times so that we have five samples of 10 flips. Suppose that the five flippers yield the following results:

Flipper 1: H T H H T T T T T T (3 H, 7 T)

Flipper 2: T T T H T H T T T T (2 H, 8 T)

Flipper 3: H T H H T H T H T T (5 H, 5 T)

Flipper 4: T T T T T T T T T H (1 H, 9 T)

Flipper 5: T H H H T H T H T H (6 H, 4 T)

If a sampling distribution of the percentage of heads is then created, the values 30, 20, 50, 10, and 60 would be included for these five samples. If we continue collecting samples of 10 flips, with enough samples we will end up with the mean of the sampling distribution equal to the true population mean of 50. However, there will be a large amount of variability between our samples, as with the five samples presented previously. If we create a sampling distribution for samples of 100 flips of a fair coin, it is extremely unlikely that the

samples will have 10% or 20% heads. In fact, we would quickly observe that although the mean of the sample statistics will be equal to the population mean of 50% heads, the sample statistics will vary much less than did the statistics for samples of 10 flips.

### Mathematical Formula

The mathematical proof that Bernoulli originally solved yields the simple formula,

$$\overline{X}_n \rightarrow \mu \text{ as } n \rightarrow \infty.$$

Mathematicians sometimes denote two versions of the Law of Large Numbers, which are referred to as the weak version and the strong version. Simply put, the weak version suggests that $\overline{X}_n$ converges in probability to $\mu$, whereas the strong version suggests that $\overline{X}_n$ converges almost surely to $\mu$.

### When Bigger Samples Are Not Better

Although larger samples better represent the populations from which they are drawn, there are instances when a large sample might not provide the best parameter estimates because it is not a "good" sample. Biased samples that are not randomly drawn from the population might provide worse estimates than smaller, randomly drawn samples. The Law of Large Numbers applies only to randomly drawn samples, that is, samples in which all members of the population have an equal chance of being selected.

*Jill H. Lohmeier*

***See also*** Central Limit Theorem; Expected Value; Random Sampling; Sample Size; Sampling Distributions; Standard Error of the Mean

#### Further Readings

Dinov, I. D., Christou, N., & Gould, R. (2009). Law of large numbers: The theory, applications, and technology-based education. *Journal of Statistics Education, 17,* 1–15.

Durrett, R. (1995). *Probability: Theory and examples.* Pacific Grove, CA: Duxbury Press.

Ferguson, T. S. (1996). *A course in large sample theory.* New York: Chapman & Hall.

Hald, A. (2007). *A history of parametric statistical inference from Bernoulli to Fisher.* New York: Springer-Verlag.

## Least Squares, Methods of

The least-squares method (LSM) is widely used to find or estimate the numerical values of the parameters to fit a function to a set of data and to characterize the statistical properties of estimates. It is probably the most popular technique in statistics for several reasons. First, most common estimators can be cast within this framework. For example, the mean of a distribution is the value that minimizes the sum of squared deviations of the scores. Second, using squares makes LSM mathematically very tractable because the Pythagorean theorem indicates that, when the error is independent of an estimated quantity, one can add the *squared* error and the *squared* estimated quantity. Third, the mathematical tools and algorithms involved in LSM (derivatives, eigendecomposition, and singular value decomposition) have been well studied for a long time.

LSM is one of the oldest techniques of modern statistics, and even though ancestors of LSM can be traced back to Greek mathematics, the first modern precursor is probably Galileo. The modern approach was first exposed in 1805 by the French mathematician Adrien-Marie Legendre in a now classic memoir, but this method is somewhat older because it turned out that, after the publication of Legendre's memoir, Carl Friedrich Gauss (the famous German mathematician) contested Legendre's priority. Gauss often did not publish ideas when he thought that they could be controversial or not yet ripe, but he would mention his discoveries when others would publish them (the way he did, for example for the discovery of non-Euclidean geometry). And in 1809, Gauss published another memoir in which he mentioned that he had previously discovered LSM and used it as early as 1795 in estimating the orbit of an asteroid. A somewhat bitter anteriority dispute followed (a bit reminiscent of the Leibniz-Newton controversy about the invention of calculus),

which however, did not diminish the popularity of this technique.

The use of LSM in a modern statistical framework can be traced to Sir Francis Galton who used it in his work on the heritability of size, which laid down the foundations of correlation and (also gave the name to) regression analysis. The two antagonistic giants of statistics Karl Pearson and Ronald Fisher, who did so much in the early development of statistics, used and developed it in different contexts (factor analysis for Pearson and experimental design for Fisher).

Nowadays, the LSM exists with several variations: Its simpler version is called ordinary least squares (OLS), and a more sophisticated version is called weighted least squares (WLS), which often performs better than OLS because it can modulate the importance of each observation in the final solution. Recent variations of the least square method are alternating least squares (ALS) and partial least squares (PLS).

## Functional Fit Example: Regression

The oldest (and still the most frequent) use of OLS was linear regression, which corresponds to the problem of finding a line (or curve) that best fits a set of data points. In the standard formulation, a set of $N$ pairs of observations $\{Y_i, X_i\}$ is used to find a function relating the value of the dependent variable ($Y$) to the values of an independent variable ($X$). With one variable and a linear function, the prediction is given by the following equation:

$$\hat{Y} = a + bX. \tag{1}$$

This equation involves two free parameters that specify the intercept ($a$) and the slope ($b$) of the regression line. The least-squares method defines the estimate of these parameters as the values that minimize the sum of the squares (hence the name *least squares*) between the measurements and the model (i.e., the predicted values). This amounts to minimizing the expression:

$$\mathcal{E} = \sum_i (Y_i - \hat{Y}_i)^2 = \sum_i [Y_i - (a + bX_i)]^2, \tag{2}$$

where $\mathcal{E}$ stands for "error," which is the quantity to be minimized. The estimation of the parameters is obtained using basic results from calculus and, specifically, uses the property that a quadratic expression reaches its minimum value when its derivatives vanish. Taking the derivative of $\mathcal{E}$ with respect to $a$ and $b$ and setting them to zero gives the following set of equations (called the *normal* equations):

$$\frac{\partial \mathcal{E}}{\partial a} = 2Na + 2b \sum X_i - 2 \sum Y_i = 0 \tag{3}$$

and

$$\frac{\partial \mathcal{E}}{\partial b} = 2b \sum X_i^2 + 2a \sum X_i - 2 \sum Y_i X_i = 0. \tag{4}$$

Solving the normal equations gives the following least-squares estimates of $a$ and $b$ as:

$$a = M_Y - bM_X \tag{5}$$

with $M_y$ and $M_X$ denoting the means of $X$ and $Y$, and

$$b = \frac{\sum (Y_i - M_Y)(X_i - M_X)}{\sum (X_i - M_X)^2}. \tag{6}$$

OLS can be extended to more than one independent variable (using matrix algebra) and to nonlinear functions.

### *The Geometry of Least Squares*

OLS can be interpreted in a geometrical framework as an orthogonal projection of the data vector onto the space defined by the independent variable. The projection is orthogonal because the predicted values and the actual values are uncorrelated. This is illustrated in Figure 1, which depicts the case of two independent variables (vectors $\mathbf{X}_1$ and $\mathbf{X}_2$) and the data vector ($\mathbf{y}$), and it shows that the error vector ($\mathbf{y} - \hat{\mathbf{y}}$) is orthogonal to the least-squares ($\hat{\mathbf{y}}$) estimate, which lies in the subspace defined by the two independent variables.

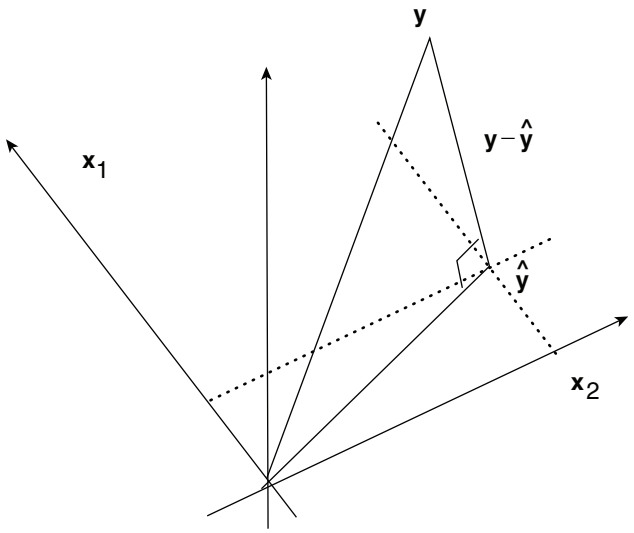



**Figure 1** The Least-Squares Estimate of the Data Is the Orthogonal Projection of the Data Vector Onto the Independent Variable Subspace

### *Optimality of Least-Squares Estimates*

OLS estimates have some strong statistical properties. Specifically when (a) the data obtained constitute a random sample from a well-defined population, (b) the population model is linear, (c) the error has a zero expected value, (d) the independent variables are linearly independent, and (e) the error is normally distributed and uncorrelated with the independent variables (the so-called homoscedasticity assumption), the OLS estimate is the *b*est *l*inear *u*nbiased *e*stimate, often denoted with the acronym "BLUE" (the five conditions and the proof are called the Gauss-Markov conditions and theorem). In addition, when the Gauss-Markov conditions hold, OLS estimates are also maximum-likelihood estimates.

### *Weighted Least Squares*

The optimality of OLS relies heavily on the homoscedasticity assumption. When the data come from different subpopulations for which an independent estimate of the error variance is available, a better estimate than OLS can be obtained using weighted least squares (WLS), which is also called generalized least squares (GLS). The idea is to assign to each observation a weight that reflects the uncertainty of the measurement. In general, the weight $w_i$, which is assigned to the $i$th observation, will be a function of the variance of this observation, which is denoted $\sigma_i^2$. A straightforward weighting schema is to define $w_i = \sigma_i^{-1}$ (but other more sophisticated weighted schemes can also be proposed). For the linear regression example, WLS will find the values of $a$ and $b$ minimizing:

$$\mathcal{E}_w = \sum_i w_i (Y_i - \hat{Y}_i)^2 = \sum_i w_i [Y_i - (a + bX_i)]^2. \quad (7)$$

### Iterative Methods: Gradient Descent

When estimating the parameters of a nonlinear function with OLS or WLS, the standard approach using derivatives is not always possible. In this case, iterative methods are often used. These methods search in a stepwise fashion for the best values of the estimate. Often they proceed by using at each step a linear approximation of the function and refine this approximation by successive corrections. The techniques involved are known as gradient descent and Gauss-Newton approximations. They correspond to nonlinear least squares approximation in numerical analysis and nonlinear regression in statistics. Neural networks constitutes a popular recent application of these techniques

## Problems with Least Squares and Alternatives

Despite its popularity and versatility, LSM has its problems. Probably the most important drawback of LSM is its high sensitivity to outliers (i.e., extreme observations). This is a consequence of using squares because squaring exaggerates the magnitude of differences (e.g., the difference between 20 and 10 is equal to 10, but the difference between $20^2$ and $10^2$ is equal to 300) and therefore gives a much stronger importance to extreme observations. This problem is addressed by using robust techniques that are less sensitive to the effect of outliers. This field is currently under development and is likely to become more important in the future.

*Hervé Abdi*

*See also* Bivariate Regression; Correlation; Multiple Regression; Pearson Product-Moment Correlation Coefficient;

**Further Reading**

Abdi, H., Valentin, D., & Edelman, B. E. (1999). *Neural networks*. Thousand Oaks, CA: Sage.
Bates, D. M., & Watts, D. G. (1988). *Nonlinear regression analysis and its applications*. New York: John Wiley.
Greene, W. H. (2002). *Econometric analysis*. New York: Prentice Hall.
Harper, H. L. (1974). The method of least squares and some alternatives. Part I, II, III, IV, V, VI. *International Statistical Review, 42,* 147–174; 235–264.
Harper, H. L. (1975). The method of least squares and some alternatives. Part I, II, III, IV, V, VI. *International Statistical Review, 43,* 1–44; 125–190; 269–272.
Harper, H. L. (1976). The method of least squares and some alternatives. Part I, II, III, IV, V, VI. *International Statistical Review, 44,* 113–159.
Nocedal, J., & Wright, S. (1999). *Numerical optimization*. New York: Springer-Verlag.
Plackett, R. L. (1972). The discovery of the method of least squares. *Biometrika, 59,* 239–251.
Seal, H. L. (1967). The historical development of the Gauss linear model. *Biometrika, 54,* 1–23.

# Levels of Measurement

How things are measured is of great importance, because the method used for measuring the qualities of a variable gives researchers information about how one should be interpreting those measurements. Similarly, the precision or accuracy of the measurement used can lead to differing outcomes of research findings, and it could potentially limit the statistical analyses that could be performed on the data collected.

Measurement is generally described as the assignment of numbers or labels to qualities of a variable or outcome by following a set of rules. There are a few important items to note in this definition. First, measurement is described as an assignment because the researcher decides what values to assign to each quality. For instance, on a football team, the coach might assign each team member a number. The actual number assigned does not necessarily have any significance, as player #12 could just have easily been assigned #20 instead. The important point is that each player was assigned a number. Second, it is also important to notice that the number or label is assigned to a quality of the variable or outcome. Each thing that is measured generally measures only one aspect of that variable. So one could measure an individual's weight, height, intelligence, or shoe size, and one would discover potentially important information about an aspect of that individual. However, just knowing a person's shoe size does not tell everything there is to know about that individual. Only one piece of the puzzle is known. Finally, it is important to note that the numbers or labels are not assigned willy-nilly but rather according to a set of rules. Following these rules keeps the assignments constant, and it allows other researchers to feel confident that their variables are measured using a similar scale to other researchers, which makes the measurements of the same qualities of variables comparable.

These scales (or levels) of measurement were first introduced by Stanley Stevens in 1946. As a psychologist who had been debating with other scientists and mathematicians on the subject of measurement, he proposed what is referred to today as the levels of measurement to bring all interested parties to an agreement. Stevens wanted researchers to recognize that different varieties of measurement exist and that types of measurement fall into four proposed classes. He selected the four levels through determining what was required to measure each level as well as what statistical processes could reasonably be performed with variables measured at those levels. Although much debate has ensued on the acceptable statistical processes (which are explored later), the four levels of measurement have essentially remained the same since their proposal so many years ago.

## The Four Levels of Measurement

### *Nominal*

The first level of measurement is called nominal. Nominal-level measurements are names or category labels. The name of the level, nominal, is said to derive from the word *nomin-*, which is a Latin prefix meaning name. This fits the level very well, as the goal of the first level of measurement is

to assign classifications or names to qualities of variables. If "type of fruit" was the variable of interest, the labels assigned might be bananas, apples, pears, and so on. If numbers are used as labels, they are significant only in that their numbers are different but not in amount. For example, for the variable of gender, one might code males = 1 and females = 2. This does not signify that there are more females than males, or that females have more of any given quality than males. The numbers assigned as labels have no inherent meaning at the nominal level. Every individual or item that has been assigned the same label is treated as if they are equivalent, even if they might differ on other variables. Note also from the previous examples that the categories at the nominal level of measurement are discrete, which means mutually exclusive. A variable cannot be both male and female in this example, only one or the other, much as one cannot be both an apple and a banana. The categories must not only be discrete, but also they must be exhaustive. That is, all participants must fit into one (and only one) category. If participants do not fit into one of the existing categories, then a new category must be created for them. Nominal-level measurements are the least precise level of measurement and as such, tell us the least about the variable being measured. If two items are measured on a nominal scale, then it would be possible to determine whether they are the same (do they have the same label?) or different (do they have different labels?), but it would not possible to identify whether one is different from the other in any quantitative way. Nominal-level measurements are used primarily for the purposes of classification.

### *Ordinal*

The second level of measurement is called ordinal. Ordinal-level measurements are in some form of order. The name, ordinal, is said to derive from the word *ordin-,* which is a Latin prefix meaning order. The purpose of this second level of measurement is to rank the size or magnitude of the qualities of the variables. For example, the order of finish of the Kentucky Derby might be Big Brown #1, Eight Belles #2, and Denis of Cork #3. In the ordinal level of measurement, there is not only category information, as in the nominal level, but also rank information. At this level, it is known that the participants are different and which participant is better (or worse) than another participant. Ordinal measurements convey information about order but still do not speak to amount. It is not possible to determine how much better or worse participants are at this level. So, Big Brown might have come in first, the second place finisher came in 4 3/4 lengths behind, and the third place finisher came in 3 1/2 lengths behind the second place finisher. However, because the time between finishers is different, one cannot determine from the rankings alone (1st, 2nd, and 3rd) how much faster one horse is than another horse. Because the differences between rankings do not have a constant meaning at this level of measurement, researchers might determine that one participant is greater than another but not how much greater he or she is. Ordinal measurements are often used in educational research when examining percentile ranks or when using Likert scales, which are commonly used for measuring opinions and beliefs on what is usually a 5-point scale.

### *Interval*

The third level of measurement is called interval. Interval-level measurements are created with each interval exactly the same distance apart. The name, interval, is said to derive from the words *inter-,* which is a Latin prefix meaning between, and *vallum,* which is a Latin word meaning ramparts. The purpose of this third level of measurement is to allow researchers to compare how much greater participants are than each other. For example, a hot day that is measured at 96 °F is 20° hotter than a cooler day measured at 76 °F, and that is the same distance as an increase in temperature from 43 °F to 63 °F. Now, it is possible to say that the first day is hotter than the second day and also how much hotter it is. Interval is the lowest level of measurement that allows one to talk about amount. A piece of information that the interval scale does not provide to researchers is a true zero point. On an interval-level scale, whereas there might be a marking for zero, it is just a place holder. On the Fahrenheit scale, zero degrees does not mean that there is no heat, because it is possible to measure negative degrees. Because of the lack of an absolute zero at this level of measurement, although it is determinable how much

greater one score is from another, one cannot determine whether a score is twice as big as another. So if Jane scores a 6 on a vocabulary test, and John scores a 2, it does not mean that Jane knows 3 times as much as John, because the zero point on the test is not a true zero, but rather an arbitrary one. Similarly, a zero on the test does not mean that the person being tested has zero vocabulary ability. There is some controversy as to whether the variables that we measure for aptitude, intelligence, achievement, and other popular educational tests are measured at the interval or ordinal levels.

### *Ratio*

The fourth level of measurement is called ratio. Ratio-level measurements are unique among all the other levels of measurements because they have an absolute zero. The name ratio is said to derive from the Latin word *ratio,* meaning calculation. The purpose of this last level of measurement is to allow researchers to discuss not only differences in magnitude but also ratios of magnitude. For example, for the ratio level variable of weight, one can say that an object that weights 40 pounds is twice as heavy as an object that weights 20 pounds. This level of measurement is so precise, it can be difficult to find variables that can be measured using this level. To use the ratio level, the variable of interest must have a true zero. Many social science and education variables cannot be measured at this level because they simply do not have an absolute zero. It is fairly impossible to have zero self-esteem, zero intelligence, or zero spelling ability, and as such, none of those variables can be measured on a ratio level. In the hard sciences, more variables are measurable at this level. For example, it is possible to have no weight, no length, or no time left. Similar to interval level scales, very few education and social science variables are measured at the ratio level. One of the few common variables at this measure is reaction time, which is the amount of time that passes between when a stimulus happens and when a reaction to it is noted. Other common occurrences of variables that are measured at this level happen when the variable is measured by counting. So the number of errors, number of items correct, number of cars in a parking lot, and number of socks in a drawer are all measured at the ratio level, because it is possible have a true zero of each (no socks, no cars, no items correct, and no errors).

## Commonalities in the Levels of Measurement

Even as each level of measurement is defined by following certain rules, the levels of measurement as a whole also have certain rules that must be followed. All possible variables or outcomes can be measured at least one level of measurement. Levels of measurement are presented in an order from least precise (and therefore least descriptive) to most precise (and most descriptive). Within this order, each level of measurement follows all of the rules of the levels that preceded it. So, whereas the nominal level of measurement only labels categories, all the levels that follow also have the ability to label categories. Each subsequent level retains all the abilities of the level that came before. Also, each level of measurement is more precise than the ones before, so the interval level of measurement is more exact in what it can measure than the ordinal or nominal levels of measurement. Researchers generally believe that any outcome or variable should be measured at the most precise level possible, so in the case of a variable that could be measured at more than one level of measurement, it would be more desirable to measure it at the highest level possible. For example, one could measure the weight of items in a nominal scale, assigning the first item as "1," the second item as "2," and so on. Or, one could measure the same items on an ordinal scale, assigning the labels of "light" and "heavy" to the different items. One could also measure weight on an interval scale, where one might set zero at the average weight, and items would be labeled based on how their weights differed from the average. Finally, one could measure each item's weight, where zero means no weight at all, as weight is normally measured on a ratio scale. Using the highest level of measurement provides researchers with the most precise information about the actual quality of interest, weight.

## What Can Researchers Do With Different Levels of Measurement?

Levels of measurement tend to be treated flexibly by researchers. Some researchers believe that there

are specific statistical analyses that can only be done at higher levels of measurement, whereas others feel that the level of measurement of a variable has no effect on the allowable statistics that can be performed. Researchers who believe in the statistical limitations of certain levels of measurement might also be fuzzy on the lines between the levels. For example, many students learn about a level of measurement called quasi-interval, on which variables are measured on an ordinal scale but treated as if they were measured at the interval scale for the purposes of statistical analysis. As many education and social science tests collect data on an ordinal scale, and because using an ordinal scale might limit the statistical analyses one could perform, many researchers prefer to treat the data from those tests as if they were measured on an interval scale, so that more advanced statistical analyses can be done.

## What Statistics Are Appropriate?

Along with setting up the levels of measurement as they are currently known, Stevens also suggested appropriate statistics that should be permitted to be performed at each level of measurement. Since that time, as new statistical procedures have been developed, this list has changed and expanded. The appropriateness of some of these procedures is still under debate, so one should examine the assumptions of each statistical analysis carefully before conducting it on data of any level of measurement.

### *Nominal*

When variables are measured at the nominal level, one might count the number of individuals or items that are classified under each label. One might also calculate central tendency in the form of the mode. Another common calculation that can be performed is the chi-square correlation, which is otherwise known as the contingency correlation. Some more qualitative analyses might also be performed with this level of data.

### *Ordinal*

For variables measured at the ordinal level, all the calculation of the previous level, nominal, might be performed. In addition, percentiles might be calculated, although with caution, as some methods used for the calculation of percentiles assume the variables are measured at the interval level. The median might be calculated as a measure of central tendency. Quartiles might be calculated, and some additional nonparametric statistics might be used. For example, Spearman's rank-order correlation might be used to calculate the correlation between two variables measured at the ordinal level.

### *Interval*

At the interval level of measurement, almost every statistical tool becomes available. All the previously allowed tools might be used, as well as many that can only be properly used starting at this level of measurement. The mean and standard deviation, both frequently used calculations of central tendency and variability, respectively, become available for use at this level. The only statistical tools that should not be used at this level are those that require the use of ratios, such as the coefficient of variation.

### *Ratio*

All statistical tools are available for data at this level.

## Controversy

As happens in many cases, once someone codifies a set or rules or procedures, others proceed to put forward statements about why those rules are incorrect. In this instance, Stevens approached his proposed scales of measurement with this idea that certain statistics should be allowed to be performed only on variables that had been measured at certain levels of measurement, much as has been discussed previously. This point of view has been called measurement directed, which means that the level of measurement used should guide the researcher as to which statistical analysis is appropriate. On the other side of the debate are researchers who identify as measurement independent. These researchers believe that it is possible to conduct any type of statistical analysis, regardless of the variable's level of measurement. An

oft-repeated statement used by measurement-independent researchers was coined by F. M. Lord, who stated, "the numbers do not know where they come from" (p. 751). This means regardless of what level of measurement was used to assign those numbers, given any set of numbers it is possible to perform any statistical calculation of interest with them.

Researchers who agree with the measurement-directed position tend to take the stance that although it is possible to conduct any statistical analysis the researcher wishes with any numbers, regardless of the level on which they were measured, it is most difficult to interpret the results of those analyses in an understandable manner. For example, if one measures gender on a nominal scale, assigning the labels "1" to males and "2" to females, one could use those numbers to calculate the mean gender in the U.S. population as 1.51. But how does one interpret that? Does it mean that the average person in the United States is male plus half a male? The average person is three quarters female? Researchers who believe in measurement-directed statistics would argue that the mode should be used to calculate central tendency for anything measured at the nominal level, because using the mode would bring more interpretable results than using the mean, which is more appropriate for interval-level data. Researchers who take the measurement-independent view believe that by restricting the statistical analyses one can conduct, one loses the use of important statistical tools that in some cases (particularly in the use of quasi-interval data) could have lead to valuable research breakthroughs.

*Carol A. Carman*

***See also*** Central Tendency, Measures of; Descriptive Statistics; Mean; Median; Mode; "On the Theory of Scales of Measurement"; Planning Research; Sensitivity; Standard Deviation; Variability, Measure of

**Further Readings**

Coladarci, T., Cobb, C. D., Minium, E. W., & Clarke, R. C. (2004). *Fundamentals of statistical reasoning in education*. Hoboken, NJ: John Wiley.

Gaito, J. (1980). Measurement scales and statistics: Resurgence of an old misconception. *Psychological Bulletin, 87,* 564–567.

Lord, F. M. (1953). On the statistical treatment of football numbers. *American Psychologist, 8,* 750–751.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Lawrence Erlbaum.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103,* 677–680.

Stine, W. W. (1989). Meaningful inference: The role of measurement in statistics. *Psychological Bulletin, 105,* 147–155.

Thorndike, R. M. (2005). *Measurement and evaluation in psychology and education* (7th ed.). Upper Saddle River, NJ: Pearson Education.

## Likelihood Ratio Statistic

The likelihood ratio statistic evaluates the relative plausibility of two competing hypotheses on the basis of a collection of sample data. The favored hypothesis is determined by whether the ratio is greater than or less than one.

To introduce the likelihood ratio, suppose that $y_{OBS}$ denotes a vector of observed data. Assume that a parametric joint density is postulated for the random vector $Y$ corresponding to the realization $y_{OBS}$. Let $f(y;\theta)$ represent this density, with parameter vector $\theta$. The *likelihood* of $\theta$ based on the data $y_{OBS}$ is defined as the joint density:

$$L(\theta; y_{OBS}) = f(y_{OBS}; \theta).$$

Although the likelihood and the density are the same function, they are viewed differently: The density $f(y;\theta)$ assigns probabilities to various outcomes for the random vector $Y$ based on a fixed value of $\theta$, whereas the likelihood $L(\theta; y_{OBS})$ reflects the plausibility of various values for $\theta$ based on the observed data $y_{OBS}$.

In formulating the likelihood, multiplicative factors that do not depend on $\theta$ are routinely omitted, and the function is redefined based on the remaining terms, which comprise the *kernel*. For instance, when considering a binomial experiment based on $n$ trails with success probability $\theta$, the density for the success count $Y$ is

$$f(y;\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y},$$

whereas the corresponding likelihood for $\theta$ based on the observed success count $y_{OBS}$ is often defined as

$$L(\theta; y_{OBS}) = \theta^{y_{OBS}}(1-\theta)^{n-y_{OBS}}.$$

Now, consider two possible values for the parameter vector $\theta$, say $\theta_0$ and $\theta_1$. The likelihood ratio statistic $L(\theta_0; y_{OBS})/L(\theta_1; y_{OBS})$ might be used to determine which of these two candidate values is more "likely" (i.e., which is better supported by the data $y_{OBS}$). If the ratio is less than one, $\theta_1$ is favored; if the ratio is greater than one, $\theta_0$ is preferred.

As an example, in a classroom experiment to illustrate simple Mendelian genetic traits, suppose that a student is provided with 20 seedlings that might flower either white or red. She is told to plant these seedlings and to record the colors after germination. Let $\theta$ denote the probability of a seedling flowering white. If $Y$ denotes the number of seedlings among the 20 planted that flower white, then $Y$ might be viewed as arising from a binomial distribution with density

$$f(y; \theta) = \binom{20}{y}\theta^{y}(1-\theta)^{n-y}.$$

The student is told that $\theta$ is either $\theta_0 = 0.75$ or $\theta_1 = 0.50$; she must use the outcome of her experiment to determine the correct probability. After planting the 20 seedlings, she observes that $y_{OBS} = 13$ flower white. In this setting, the likelihood ratio statistic

$$L(\theta_0; y_{OBS})/L(\theta_1; y_{OBS})$$

equals 1.52. Thus, the likelihood ratio implies that the value $\theta_0 = 0.75$ is the more plausible value for the probability $\theta$. Based on the ratio, the student should choose the value $\theta_0 = 0.75$.

The likelihood ratio might also be used to test formally two competing point hypotheses $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$. In fact, the Neyman–Pearson Lemma establishes that the power of such a test will be at least as high as the power of any alternative test, assuming that the tests are conducted using the same levels of significance.

A generalization of the preceding test allows one to evaluate two competing composite hypotheses $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$. Here, $\Theta_0$ and $\Theta_1$ refer to disjoint parameter spaces where the parameter vector $\theta$ might lie. The conventional test statistic, which is often called the *generalized likelihood ratio statistic,* is given by

$$L(\hat{\theta}_0; y_{OBS})/L(\hat{\theta}; y_{OBS}),$$

where $L(\hat{\theta}_0; y_{OBS})$ denotes the maximum value attained by the likelihood $L(\theta; y_{OBS})$ as the parameter vector $\theta$ varies over the space $\Theta_0$, and $L(\hat{\theta}; y_{OBS})$ represents the maximum value attained by $L(\theta; y_{OBS})$ as $\theta$ varies over the combined space $\Theta_0 \cup \Theta_1$.

Tests based on the generalized likelihood ratio are often optimal in terms of power. The size of a test refers to the level of significance at which the test is conducted. A test is called uniformly most powerful (UMP) when it achieves a power that is greater than or equal to the power of any alternative test of comparable size. When no UMP test exists, it might be helpful to restrict attention to only those tests that can be classified as unbiased. A test is unbiased when the power of the test never falls below its size [i.e., when $\Pr(\text{reject } H_0 | \theta \in \Theta_1) \geq \Pr(\text{reject } H_0 | \theta \in \Theta_0)$]. A test is called uniformly most powerful unbiased (UMPU) when it achieves a power that is greater than or equal to the power of any alternative unbiased test. The generalized likelihood ratio statistic can often be used to formulate UMP and UMPU tests.

The reliance of the likelihood ratio statistic in statistical inference is largely based on the *likelihood principle*. Informally, this principle states that all the information in the sample $y_{OBS}$ that is relevant for inferences on the parameter vector $\theta$ is contained within the likelihood function $L(\theta; y_{OBS})$. The likelihood principle is somewhat controversial and is not universally held. For instance, neglecting constants that do not involve $\theta$ the same likelihood might result from two different experimental designs. In such instances, likelihood-based inferences would be the same under either design, although tests that incorporate the nature of the design might lead to different conclusions. For instance, consider the preceding genetics example based on simple Mendelian traits. If a student were to plant all $n$ seedlings at one time and to count the number $Y$ that eventually flower white, then the count $Y$ would follow a binomial distribution. However, if the student were to plant

seedlings consecutively one at a time and continue until a prespecified number of seedlings flower red, then the number $Y$ that flower white would follow a negative binomial distribution. Based on the kernel, each experimental design leads to the same likelihood. Thus, if the overall number of seedlings planted $n$ and the observed number of white flowering seedlings $y_{OBS}$ are the same in each design, then likelihood-based inferences such as the preceding likelihood ratio test would yield identical results. However, tests based on the probability distribution models $f(y;\theta)$ could yield different conclusions.

The likelihood ratio $L(\theta_0; y_{OBS})/L(\theta_1; y_{OBS})$ has a simple Bayesian interpretation. Prior to the collection of data, suppose that the candidate values $\theta_0$ and $\theta_1$ are deemed equally likely, so that the prior probabilities $\Pr(\theta_0) = \Pr(\theta_1) = 0.5$ are employed. By Bayes's rule, the ratio of the posterior probabilities for the two parameter values,

$$\Pr(\theta_0|y_{OBS})/\Pr(\theta_1|y_{OBS}),$$

corresponds to the likelihood ratio. As this interpretation would suggest, the concept of the likelihood function and the likelihood principle both play prominent roles in Bayesian inference.

*Joseph E. Cavanaugh and Eric D. Foster*

***See also*** Bayes's Theorem; Directional Hypothesis; Hypothesis; Power; Significance Level, Concept of

**Further Readings**

Pawitan, Y. (2001). *In all likelihood: Statistical modeling and inference using likelihood.* Oxford, UK: Oxford University Press.

# Likert Scaling

Likert (pronounced lick-ert) scaling is a method of attitude, opinion, or perception assessment of a unidimensional variable or a construct made up of multidimensions or subscales. It recognizes the contribution to attitude assessment of Rensus Likert who published a classic paper on this topic in 1932, based on his doctoral dissertation directed by Gardner Murphy and based on work Murphy had undertaken in 1929. The use of Likert items and scaling is probably the most used survey methodology in educational and social science research and evaluation.

The Likert scale provides a score based on a series of items that have two parts. One part is the stem that is a statement of fact or opinion to which the respondent is asked to react. The other part is the response scale. Likert was the first recognized for the use of a 5-point, ordinal scale of *strongly approve—approve—undecided—disapprove—strongly disapprove*. The scale is often changed to other response patterns such as *strongly agree—agree—neutral—disagree—strongly disagree*. This entry discusses Likert's approach and scoring methodology and examines the research conducted on Likert scaling and its modifications.

## Likert's Approach

In Likert's original research, which led to Likert scaling, Likert compared four ways of structuring attitude survey items believing that there was an alternative to the approach attributed to Louis Leon Thurstone. Although both approaches were based on equal-interval, ordinal stepped scale points, Likert considered Thurstone's methods to be a great deal of work that was not necessary. Setting up a Thurstone scale involved the use of judges to evaluate statements to be included in the survey. This included rank ordering the statements in terms of the expected degree of the attribute being assessed and then comparing and ordering each pair of item possibilities, which is an onerous task if there were many item possibilities. Originally, each item was scored as a dichotomy (agree/disagree or +/−).

A Thurstone scale was scored in a similar manner as Likert's original method using sigma values, which were $z$ scores weighted by the responses corresponding to the assumed equal interval categories. However, part of the problem with Thurstone's scoring method related to having a spread of judge-determined 1 to 11 scoring categories when scoring the extreme values of 0 or 1 proportions. These could not be adequately accounted for because they were considered as $\pm$ infinity $z$ values in a sigma scoring approach and thus were dropped from the scoring. Likert felt

there was another approach that did not rely so much on the use of judges and could include the scoring of items where everyone either did not or did select the extreme score category by using the $\pm 3.00$ $z$ values instead of $\pm \infty$. Thus, Likert set out to use some of the features of a Thurstone scale but simplify the process and hope to achieve a similar level of reliability found with the Thurstone scale. His research met the goals he set out to meet.

A stem or statement was presented related to racial attitudes and then respondents were asked to respond to one of several response sets. One set used "yes / no" options, and another used narrative statements, and two of them used what we now know as a Likert item, using *strongly approve—approve—undecided—disapprove—strongly disapprove* as the response categories. The distinction between the last two types of items relates to the source of the questions as being developed specifically by Likert for assessing attitudes and the other as abbreviations of newspaper articles reflecting societal conflicts among race-based groups.

## Likert's Scoring Methodology

Likert found that many items he used had distributions resembling a normal distribution. He concluded that if these distributions resembled a normal distribution, it was legitimate to determine a single unidimensional scale value by finding the mean or sum of the items and using that for a value that represented the attitude, opinion, or perception of the variable on a continuum.

Sigma values were $z$ scores weighted by the use of responses to the five categories. These were then used by item to estimate score reliabilities (using split-half and test–retest approaches), which were found to be high. Likert also demonstrated a high level of concurrent validity between his approach and Thurstone's approach, even though he had used only about half the number of items that Thurstone had used.

Likert sings the praises of the sigma scoring technique. However, he also discovered that simply summing the scores resulted in about the same degree of score reliability, for both split-half and test–retest score reliabilities, as the sigma approach. Thus was born the concept of Likert scaling involving the use of the mean or sum of scores from a set of items to represent a position on the attitude variable continuum.

Although Likert focused on developing a unidimensional scale, applications of his methodology have been used to develop multidimensional scales that include subscales intended to assess attitudes and opinions on different aspects of the construct of interest.

## Modifications

Many modifications of Likert scaling use a wide variety of response sets similar or not so similar to the ones Likert used that are called Likert-type items. It seems that almost any item response set that includes ordered responses in a negative and positive direction gets labeled as a Likert item. More than 35 variations of response sets have been identified, even though many of them vary considerably from Likert's original response set. Some even use figures such as smiley or frowny faces instead of narrative descriptions or abbreviations. These scales have been used often with children.

## Research

### *Controversial Issues Related to Likert Items and Scales*

Research on Likert item stem and response construction, overall survey design, methods of scoring, and various biases has been extensive; it is probably one of the most researched topics in social science. There are many controversial issues and debates about using Likert items and scales, including the reading level of respondents, item reactivity, the length or number of items, the mode of delivery, the number of responses, using an odd or even number of responses, labeling of a middle response, the direction of the response categories, dealing with missing data, the lack of attending behaviors, acquiescence bias, central tendency bias, social desirability bias, the use of parametric methods or nonparametric methods when comparing scale indicators of central tendency (median or mean), and, probably most controversial, the use of negatively worded items. All of these have the potential for influencing score reliability and

validity, some more than others, and a few actually increase the estimate of reliability.

Because Likert surveys are usually in the category of "self-administered" surveys, the reading level of respondents must be considered. Typically, a reading level of at least 5th grade is often considered a minimal reading level for surveys given to most adults in the general population. Often, Edward Fry's formula is used to assess reading level of a survey. A companion issue is when surveys are translated from one language to another. This can be a challenging activity that, if not done well, can reduce score reliability. Related to this is the potential for reducing reliability and validity when items are reactive or stir up emotions in an undesirable manner that can confound the measure of the attitudes of interest. Sometimes, Likert survey items are read to respondents in cases where reading level might be an issue or clearly when a Likert scale is used in a telephone survey. Often, when reading Likert response options over the phone, it is difficult for some respondents to keep the categories in mind, especially if they change in the middle of the survey. Other common modes of delivery now include online Likert surveys used for myriad purposes.

Survey length can also affect reliability. Even though one way of increasing score reliability is to lengthen a survey, making the survey too long and causing fatigue or frustration will have the opposite effect.

One issue that often comes up is deciding on the number of response categories. Most survey researchers feel three categories might be too few and more than seven might be too many. Related to this issue is whether to include an odd or even number of response categories. Some feel that using an even number of categories forces the respondent to chose one directional opinion or the other, even if mildly so. Others feel there should be an odd number of responses and the respondent should have a neutral or nonagree or nondisagree opinion. If there are an odd number of response categories, then care must be used in defining the middle category. It should represent a point of the continuum such as *neither approve nor disapprove, neither agree nor disagree,* or *neutral.* Responses such as *does not apply* or *cannot respond* do not fit the ordinal continuum. These options can be used, but it is advisable not to put these as midpoints on an ordinal continuum or to give them a score for scaling purposes.

Another issue is the direction of the response categories. Options are to have the negative response set on the left side of the scale moving to the right becoming more positive or having the positive response set on the left becoming more negative as the scale moves from left to right. There does not seem to be much consensus on which is better, so often the negative left to positive right is preferred.

Missing data are as much an issue in Likert scaling as in all other types of research. Often, a decision needs to be made relative to how many items need to be completed for the survey to be considered viable for inclusion in the data set. Whereas there are no hard-and-fast rules for making this decision, most survey administrators would consider a survey with fewer than 80% of the items completed not to be a viable entry. There are a few ways of dealing with missing data when there are not a lot of missed responses. The most common is to use the mean of the respondent's responses on the completed items as a stand-in value. This is done automatically if the scored value is the mean of the answered responses. If the sum of items is the scale value, any missing items will need to have the mean of the answered items imputed into the missing data points before summing the items to get the scaled score.

### *Response Bias*

Several recognized biased responses can occur with Likert surveys. Acquiescence bias is the tendency of the respondent to provide positive responses to all or almost all of the items. Of course, it is hard to separate acquiescence bias from reasoned opinions for these respondents. Often, negatively worded Likert stems are used to determine whether this is happening based on the notion that if a respondent responded positively both to items worded in a positive direction as well as a negative direction, then they were more likely to be exhibiting this biased behavior rather than attending to the items. Central tendency bias is the tendency to respond to all or most of the

items with the middle response category. Using an even number of response categories is often a strategy employed to guard against this behavior. Social desirability bias is the tendency for respondents to reply to items to reflect what they believe they would be expected to respond based on societal norms or values rather than their own feelings. Likert surveys on rather personal attitudes or opinions related to behaviors considered by society to be illegal, immoral, unacceptable, or personally embarrassing are more prone to this problem. This problem is exacerbated if respondents have any feeling that their responses can be directly or even indirectly attributed to them personally. The effect of two of these behaviors on reliability is somewhat predictable. It has been demonstrated that different patterns of responses have differential effects on Cronbach's alpha coefficients. Acquiescent (or the opposite) responses inflate Cronbach's alpha. Central tendency bias has little effect on Cronbach's alpha. It is pretty much impossible to determine the effect on alpha from social desirability responses, but it would seem that there would not be a substantial effect on it.

### *Inferential Data Analysis of Likert Items and Scale Scores*

One of the most controversial issues relates to how Likert scale data can be used for inferential group comparisons. On an item level, it is pretty much understood that the level of measurement is ordinal and comparisons on an item level should be analyzed using nonparametric methods, primarily using tests based on the chi-square probability distribution. Tests for comparing frequency distributions of independent groups often use a chi-square test of independence. When comparing Likert item results for the dependent group situation such as a pretest–posttest arrangement, McNemar's test is recommended. The controversy arises when Likert scale data (either in the form of item means or sums) are being compared between or among groups. Some believe this scale value is still at best an ordinal scale and recommend the use of a nonparametric test, such as a Mann–Whitney, Wilcoxon, or Kruskal–Wallis test, or the Spearman rank-order correlation if looking for variable relationships. Others are willing to believe the assumption, which was pretty much promulgated by Likert, that item distributions are close to being normal and are thus additive, giving an approximate interval scale. This would justify the use of $z$ tests, $t$ tests, and analysis of variance for group inferential comparisons and the use of Pearson's $r$ to examine variable relationships. Rasch modeling is often used as an approach for obtaining interval scale estimates for use in inferential group comparisons if certain item characteristics are assumed. To the extent the item score distributions depart from normality, this assumption would have less viability and would tend to call for the use of nonparametric methods.

### *Use of Negatively Worded or Reverse-Worded Likert Stems*

Although there are many controversies about the use of Likert items and scales, the one that seems to be most controversial is the use of reverse or negatively worded Likert item stems. This has been a long recommended practice to guard against acquiescence. Many Likert item scholars still recommend this practice. It is interesting to note that Likert used some items with positive attitude stems and some with negative attitude stems in all four of his types of items. However, Likert provides no rationale for doing this in his classic work. Many researchers have challenged this practice as not being necessary in most attitude assessment settings and as a practice that actually reduces internal consistency score reliability. Several researchers have demonstrated that this practice can easily reduce Cronbach's alpha by at least 0.10. It has been suggested that the reversal of Likert response sets for half of the items while keeping the stems all going in a positive direction accomplishes the same purpose of using negatively worded Likert items.

### *Reliability*

Even though Likert used split-half methods for estimating score reliability, most of the time in current practice, Cronbach's alpha coefficient of internal consistency, which is also known as the Kuder-Richardson 20 approach, is used. Cronbach's alpha is sometimes defined as the mean split-half reliability coefficient if all the possible split-half coefficients are defined.

### Likert's Contribution to Research Methodology

Likert's contribution to the method of scaling, which was named for him, has had a profound effect on the assessment of opinions and attitudes of groups of individuals. There are still many issues about the design, application, scoring, and analysis of Likert scale data. However, the approach is used throughout the world, providing useful information for research and evaluation purposes.

*J. Jackson Barnette*

***See also*** Likert Scaling; "Technique for the Measurement of Attitudes, A"

#### Further Readings

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 561–573.

Babbie, E. R. (2005). *The basics of social research*. Stamford, CT: Thomson Wadsworth.

Barnette, J. J. (1999). Nonattending respondent effects on the internal consistency of self-administered surveys: A Monte Carlo simulation study. *Educational and Psychological Measurement, 59,* 38–46.

Barnette, J. J. (2000). Effects of stem and Likert response option reversals on survey internal consistency: If you feel the need, there's a better alternative to using those negatively-worded stems. *Educational and Psychological Measurement, 60,* 361–370.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.

Fry, E. (1968). A readability formula that saves time. *Journal of Reading, 11,* 265–271.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 140, 1–55.

Murphy, G. (1929). *An historical introduction to modern psychology,* New York: Harcourt.

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology, 33,* 529–554.

Uebersax, J. (2006). Likert scales: Dispelling the confusion. *Statistical Methods for Rater Agreement*. Retrieved February 16, 2009, from http://ourworld.compuserve.com/homepages/jsuebersax/likert.htm

Vagias, W. M. (2006). *Likert-type scale response anchors*. Clemson, SC: Clemson University, Clemson International Institute for Tourism & Research Development, Department of Parks, Recreation and Tourism Management.

## Line Graph

A line graph is a way of showing the relationship between two interval- or ratio-level variables. By convention, the independent variable is drawn along the abscissa ($x$-axis), and the dependent variable on the ordinate ($y$-axis). The $x$-axis can be either a continuous variable (e.g., age) or time. It is probably the most widely used type of chart because it is easy to make and the message is readily apparent to the viewer. Line graphs are not as good as tables for displaying actual values of a variable, but they are far superior in showing relationships between variables and changes over time.

### History

The idea of specifying the position of a point using two axes, each reflecting a different attribute, was introduced by René Descartes in 1637 (what are now called Cartesian coordinates). During the following century, graphs were used to display the relationship between two variables, but they were all hypothetical pictures and were not based on empirical data. William Playfair, who has been described as an "engineer, political economist, and scoundrel," is credited with inventing the line graph, pie chart, and bar chart. He first used line graphs in a book titled *The Commercial and Political Atlas,* which was published in 1786. In it, he drew 44 line and bar charts to describe financial statistics, such as England's balance of trade with other countries, its debt, and expenditures on the military.

### Types

Perhaps the most widely used version of the line graph has time along the horizontal ($x$) axis and the value of some variable on the vertical ($y$) axis. For example, it is used on weather channels to display changes in temperature over a 12- or 24-hour span, and by climatologists to show changes in average temperature over a span of centuries. The time variable can be calendar or clock time, as in

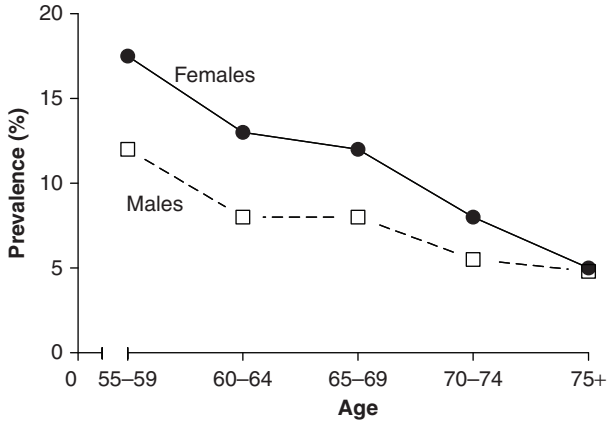



**Figure 1** Prevalence of Lifetime Mood Disorder by Age and Sex

*Source:* From "The Epidemiology of Psychological Problems in the Elderly," by D. L. Streiner, J. Cairney, & S. Veldhuizen, 2006, *Canadian Journal of Psychiatry, 51,* pp. 185–191. Copyright 1991 by the Canadian Journal of Psychiatry. Adapted and used with permission.

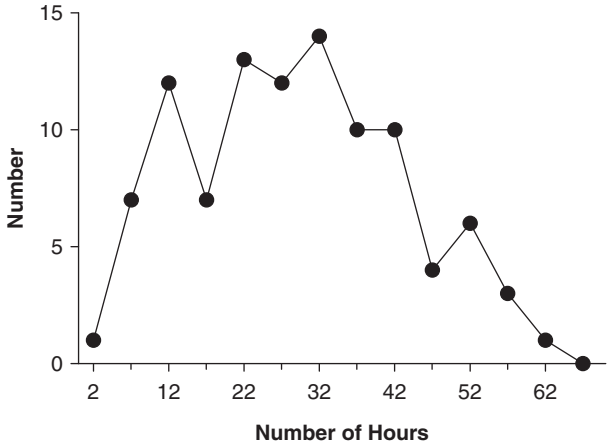



**Figure 2** Number of Hours of TV Watched Each Week by 100 People

these examples, or relative time, based on a person's age, as in Figure 1. Graphs of this latter type are used to show the expected weight of infants and children at various ages to help a pediatrician determine whether a child is growing at a normal rate. The power of this type of graph was exemplified in one displaying the prevalence of Hodgkin's lymphoma as a function of age, which showed an unusual pattern, in that there are two peaks: one between the ages of 15 to 45 and another in the mid-50s. Subsequent research, based on this observation, revealed that there are actually two subtypes of this disorder, each with a different age of onset.

Also widely used are line graphs with a continuous variable, such as weight, displayed on the abscissa, and another continuous variable (e.g., serum cholesterol) on the ordinate. As with time graphs, these allow an immediate grasp of the relationship between the variables, for example, whether they are positively or negatively correlated, whether the relationship is linear or follows some other pattern, whether it is the same or different for various groups, and so on. This type of display is extremely useful for determining whether the variables meet some of the assumptions of statistical tests, which might require, for example, a linear association between the variables.

Frequency polygons are often used as a substitute for histograms. If the values for the *x*-axis are in "bins" (e.g., ages 0–4, 5–9, 10–14, etc.), then the point is placed at the midpoint of the bin (e.g., ages 2, 7, 12, etc.), with a distance along the *y*-axis corresponding to the number or percentage of people in that category, as shown in Figure 2 (the data are fictitious). Most often, choosing between a histogram and a frequency polygon is a matter of taste; they convey identical information. The only difference is that, by convention, there are extra bins at the ends, so that the first and last values drawn are zero. Needless to say, these are omitted if the *x* values are nonsensical (e.g., an age less than zero, or fewer than 0 hours watching TV).

A useful variation of the frequency polygon is the cumulative frequency polygon. Rather than plotting the number or percentage at each value of *x*, what is shown is the cumulative total up to and including that value, as in Figure 3. If the *y*-axis shows the percentage of observations, then it is extremely easy to determine various centiles. For example, drawing a horizontal line from the *y* values of 25, 50, and 75, and dropping vertical lines to the *x*-axis from where they intersect the line yields the median and the inter-quartile range, also shown in Figure 3.

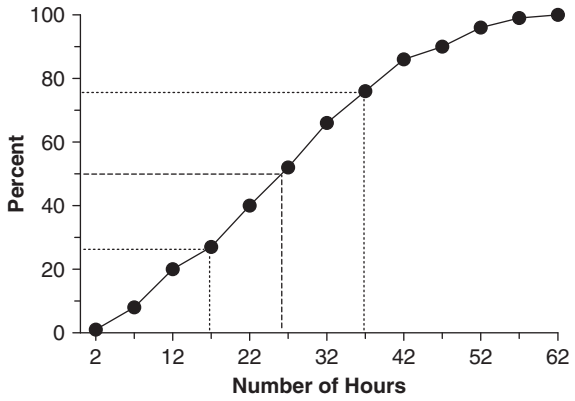



**Figure 3** Cumulative Frequency Polygon of the Data in Figure 2

## Guidelines for Drawing Line Graphs

General guidelines for creating a line graph follow.

1. Whether the *y*-axis should start at zero is a contentious issue. On the one hand, starting it at some other value has the potential to distort the picture by exaggerating small changes. For example, one advertisement for a breakfast cereal showed a decrease in eaters' cholesterol level, with the first point nearly 93% the way up the *y*-axis, and the last point, 3 weeks later, only 21% the vertical distance, which is a decrease of 77%. However, the *y*-axis began at 196 and ended at 210, so that the change of 10 points was less than 5% (both values, incidentally, are well within the normal range). On the other hand, if the plotted values are all considerably more than zero, then including zero on the axis means that 80% to 90% of the graph is blank, and any real changes might be hidden. The amount of perceived change in the graph should correspond to amount of change in the data. One way to check this is by using the formula:

$$\text{Graph Distortion Index (GDI)} = \frac{\text{Size of effect in graph}}{\text{Size of effect in data}} - 1.$$

The GDI is 0 if the graph accurately reflects the degree of change. If starting the *y*-axis at zero results in a value much more or less than 0, then the graph should start at some higher value. However, if either axis has a discontinuity in the numbering, the reader must be alerted to this by using a scale break, as in the *x*-axis of Figure 1. The two small lines (sometimes a z-shaped line is used instead) is a signal that there is a break in the numbering.

2. Nominal- or ordinal-level data should never be used for either axis. The power of the line graph is to show relationships between variables. If either variable is nominal (i.e., unordered categories), then the order of the categories is completely arbitrary, but different orderings result in different pictures, and consequently all are misleading. This is not an issue with ranks or ordered categories (i.e., ordinal data), because the ordering is fixed. However, because the spacing between values is not constant (or even known), the equal spacing along the axes gives an erroneous picture of the degree of change from one category to the next. Only interval or ratio data should be plotted with line graphs.

3. If two or more groups are plotted on the same graph, the lines should be easily distinguishable. One should be continuous, another dashed, and a third dotted, for example. If symbols are also used to indicate the specific data points, they should be large enough to be easily seen, of different shapes, and some filled and others not. Unfortunately, many graphing packages that come with computers do a poor job of drawing symbols, using an X or + that is difficult to discern, especially when the line is not horizontal. The user might have to insert better ones one manually, using special symbols.

4. Where there are two or more lines in the graph, the labels should be placed as close to the respective lines as possible, as in Figure 1. Using a legend outside the body of the graph increases the cognitive demand on the viewer, who must first differentiate among the different line and symbol types, and then shift attention to another part of the picture to find them in the legend box, and then read the label associated with it. This becomes increasingly more difficult as the number of lines increases. Boxes containing all the legends should be used only when putting the names near the appropriate lines introduces too much clutter.

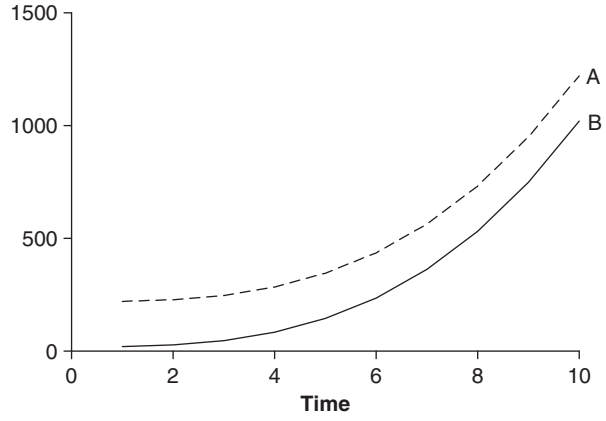



**Figure 4** Change in Two Groups Over Time

5. It is easy to determine from a line graph whether one or two groups are changing over time. It might seem, then, that it would be simple to compare the rate of change of two groups. But, this is not the case. If the lines are sloping, then it is very difficult to determine whether the difference between them is constant or is changing. For example, it seems in Figure 4 that the two groups are getting closer together over time. In fact, the difference is a constant 200 points across the entire range. The eye is fooled because the difference looks larger when the lines are horizontal than when they are nearly vertical. If the purpose is to show differences over time, it is better to plot the actual difference, rather than the individual values

6. Determining the number of labels to place on each axis is a balance between having so few that it is hard to determine where each point lies and having so many that they are crowded together. In the end, it is a matter of esthetics and judgment

*David L. Streiner*

***See also*** Graphical Display of Data; Histogram; Pie Chart

**Further Readings**

Cleveland, W. S. (1985). *The elements of graphing data.* Pacific Grove, CA: Wadsworth.

Robbins, N. B. (2005). *Creating more effective graphs.* Hoboken, NJ: John Wiley.

Streiner, D. L., & Norman, G. R. (2007). *Biostatistics: The bare essentials* (3rd ed.). Shelton, CT: PMPH.

# LISREL

LISREL, which is an acronym for linear structural relations, is a statistical program package particularly designed to estimate structural equation models (SEMs). It can also been used for several other types of analysis, such as data manipulation, exploratory data analyses, and regression, as well as factor analytic procedures. In the past few decades, SEM has become an increasingly popular technique for the analysis of nonexperimental data in the social sciences. Among programs from which researchers wishing to apply SEM might choose, such as AMOS, EQS, Mplus, SAS CALIS, and RAMONA among many others, LISREL is arguably the most longstanding and widely used tool. Notably, LISREL has been the prototype for many later developed SEM programs. After a brief history, this entry discusses the LISREL model and its execution.

## Background and Brief History

The LISREL model and computer program was developed in the 1970s by Karl G. Jöreskog and Dag Sörbom, who were both professors at Uppsala University, Sweden. In 1973, Jöreskog discovered a "maximum likelihood estimation" computational procedure and created a computer program for fitting factor models to data based on this estimation. A few years later, he, together with Sörbom, developed a program called LISREL, which incorporates maximum likelihood estimation procedures for both confirmatory factor analysis and the linear structural model among factors. To date, LISREL has undergone a few revisions. LISREL is available on a variety of operation systems such as Microsoft Windows, Macintosh, Mainframe, and UNIX. The most current version as of December 2008 is LISREL 8.8.

## The General LISREL Model

In general, LISREL estimates the unknown coefficients of a set of linear structural equations. A full LISREL model consists of two submodels: the measurement model and the structural equation

model. These models can be described by the following three equations:

1. The structural equation model: $\eta = B\eta + \Gamma\xi + \zeta$
2. The measurement model for *Y*: $y = \Lambda_y\eta + \varepsilon$
3. The measurement model for *X*: $x = \Lambda_x\xi + \delta$

### *Types of Variables*

In specifying structural equation models, one needs to be familiar with several types of variables. LISREL distinguishes variables between latent variables and observed variables. Latent variables are variables that are not observed or measured directly. They are theoretical concepts that can only be indexed by observed behaviors. Of the two types of latent variables, exogenous variables are variables that are not influenced by other variables in the model, whereas endogenous variables are the ones influenced by other variables. In other words, exogenous latent variables are independent variables and endogenous variables are dependent variables, which are influenced by the exogenous variables in the model.

### *The Measurement Model*

The measurement model (also known as the CFA model) specifies how latent variables or hypothetical constructs are indicated by the observed variables. It is designed particularly to describe the measurement properties of the observed variables. It can be specified as *X* variables or *Y* variables.

### *The Structural Equation Model*

The structural model describes the causal relations among latent variables, or how the latent variables are linked to each other. In addition, it assigns the explained and unexplained variance.

### *Greek Notation and Matrices in LISREL*

The command language of LISREL is based on a matrix representation of the confirmatory factor analysis or full structural equation model. Prior to version 8, LISREL required the use of Greek letters to specify models. Later on, LISREL used Greek as well as English, which was made possible by the SIMPLIS command language. However, familiarity with LISREL matrices and their Greek representation is helpful to master this program fully.

#### Greek Notation

Matrices are represented by uppercase Greek letters, and their elements are represented by lowercase Greek letters. The elements represent the parameters in the models. For instance, the exogenous variables are termed *X* variables, and the endogenous ones are *Y* variables.

#### Basic Matrices Within the LISREL Framework

Setting up LISREL involves specifying several matrices and specifying whether the elements within these matrices are fixed at particular values or are free parameters to be estimated by the program.

The eight matrices in the LISREL model are as follows: lambda-*X*, lambda-*Y*, theta delta, theta epsilon, psi, phi, gamma, and beta. For example, $\Lambda x$ (lambda-*X*) is a regression matrix that relates exogenous latent variables to the observed variables that are designed to measure them; $\Gamma$ (gamma) is a matrix of coefficients that relates exogenous latent variables to endogenous factors. Reviews of LISREL matrices and their Greek and program notations can be found in the LISREL manual and other SEM textbooks.

A measurement model can be defined by a regression matrix that relates the latent variable (*Y* or *X*) to its observed measures, one vector of latent variable, and one vector of measurement errors. Similarly, a structural model is defined by two matrices and three vectors.

### *Statistical Identification of the Model*

LISREL requires an establishment of the identification of the models. The issue of identification pertains to whether or not there is a unique set of parameters consistent with the data. A structural model can be just identified, overidentified, or underidentified. An overidentified model is desired in running SEMs.

#### Recommendations for Overidentification

It is recommended that latent constructs are measured by at least three measures to ensure overidentification. Recursive models with identified constructs are always identified.

## Working With LISREL

### *Input*

Execution of LISREL programs can be described by a syntax file consisting of at least three steps of operation: data input, model specification, and output of results. Each step of the operation is initiated with a command (or keyword). There are three required commands in a given LISREL input file: DA, MO, and OU. In addition, there are other commands that are useful to include, such as NI, NO, and MA.

#### Basic Rules

LISREL is controlled by two-letter keywords that represent the control lines and the names of parameters. Although a keyword might contain several letters, only the first two can be recognized by the program. Keywords are not case sensitive. In other words, they can be written in either uppercase or lowercase. However, they must be separated by blanks. Each section of the LISREL input file starts with a control line.

After optional title lines, a DA (data) command always comes first. LK, LE, FR, FI, EQ, CO, IR, PA, VA, ST, MA, PL, and NF commands must always come after the MO (model) command. The MO command must appear unless no LISREL model is analyzed.

#### Data Specification

The data specification section always starts with a DA command. To fully describe the data set, other keywords are needed to specify the number of input variables, the number of observations, and the matrix to be analyzed. Other optional input such as variable selection can be included as well.

#### Model Specification

To specify the model under study, the user needs to provide information about the number of observed variables, the number of latent variables, the form of each matrix to be analyzed, and the estimation mode (being fixed or free) of each matrix.

### *Output*

The OU command is often used to specify the methods of estimation and to specify the output to be generated. The first part of an output is a reproduction of the command file, which reminds the user of the model specification. Model specification is followed by standard errors and $t$ values together with parameter estimates. In LISREL 8 and later versions, standard errors (*SE*s) and $t$ values (TVs) are always printed, by default. These parameter estimates and other estimates derived from them determine the goodness of fit between the model under study and the observed data. In addition, modification indices (MIs) are also included by default in LISREL 8.

There are several output options from which the user can choose. All options and keywords on the OU command can be omitted; however, a line with the two letters OU must be included as the last line of the command file.

#### Path Diagrams

As part of its output, LISREL generates path diagrams that might help researchers to make sure that the model is specified correctly. Path diagrams provide a visual portrayal of relations assumed to hold among the variables in the LISREL model.

Observed variables are represented by rectangular boxes, and latent variables are represented by ellipses. Curved, double-headed arrows represent correlations between pairs of variables. A straight one-headed arrow is drawn from an exogenous variable to an endogenous variable and thus indicates the influence of one variable on another. A straight single-headed arrow is drawn to the latent variable from each of its manifest variables. The error variables appear in the diagram but are not enclosed. Variation and covariation in the dependent variables is to be accounted for or explained by the independent variables.

## Evaluation of the LISREL Model

The adequacy of the hypothesized model can be evaluated by the following indicators provided by LISREL.

### *Offending Estimates*

The appropriateness of standard errors can reflect the goodness of model fit. Excessively large or small standard errors indicate poor model fit.

Other estimates such as negative error variances and standardized coefficients exceeding 1 signal problems with model fit.

### *Overall Goodness-of-Fit Measures*

Several indexes that indicate the overall fit of a model are provided in the LISREL program, which include: chi-square, goodness-of-fit index (GFI), adjusted goodness-of-fit index (AGFI), and root mean square error of approximation (RMSEA).

### *Model Modification Indices*

Model modification indices are measures of the predicted decrease in chi-square if fixed parameters are relaxed and the model is reestimated. The fixed parameters corresponding to large modification indices are the ones that will improve model fit substantially, if freed.

## LISREL Package: LISREL, PRELIS, and SIMPLIS

### *PRELIS*

The PRELIS program is a companion program that serves as a preprocessor for LISREL, and thus the acronym PRElis. It is used for calculating correlation and covariance matrices and for estimating asymptotic covariance from raw data. It can also be used to manipulate data and to provide initial descriptive statistics and graphical displays of the data. It can prepare the correct matrix to be read by LISREL for many types of data such as continuous, ordinal, censored, or any combination thereof, even when such data are severely skewed or have missing values. In addition, PRELIS has many data management functions such as variable transformation and recoding, case selection, new variable computation, and data file merging, as well as bootstrap sampling.

A PRELIS input file is comprised of two-letter keywords that initiate and/or comprise control lines. The information in these control lines informs the program of the location of the data to be imported, what analysis to do, the destination of the newly created matrices, and what to print on the output file. DA (data), RA (raw), and OU (output) are three required control lines.

### *SIMPLIS*

SIMPLIS is a new command language that was created to simplify the use of LISREL. SIMPLIS means a simplified LISREL. With a few exceptions, any model that can be specified for use in LISREL can be specified using SIMPLIS commands. SIMPLIS commands are written in English. A SIMPLIS input file consists of the following six sections: title, observed variables, form of input data, number of cases, unobserved variables, and model structure. A variety of optional commands is also available to users.

## Statistical Applications of LISREL for Windows

The latest LISREL for Windows includes several statistical applications beyond SEM, such as follows: MULTILEV for hierarchical linear modeling, SURVEYGLIM for generalized linear modeling, CATFIRM for formative inference-based recursive modeling for categorical response variables, CONFIRM for formative inference-based recursive modeling for continuous response variables, and MAPGLIM for generalized linear modeling for multilevel data.

## Availability: Downloads and Manuals

LISREL and its companion program PRELIS are a software product marketed by Scientific Software, International. The three manuals authored by Karl G. Jöreskog and Dag Sörbom for use with LISREL and its companion package PRELIS are as follows:

*LISREL 8: Structural Equation Modeling With the SIMPLIS Command Language*

*LISREL 8: User's Reference Guide*

*PRELIS 2: User's Reference Guide*

A student edition of LISREL can be downloaded from the website of Scientific Software, International.

*Yibing Li*

***See also*** Confirmatory Factor Analysis; Endogenous Variables; Exogenous Variables; Path Analysis; Structural Equation Modeling

**Further Readings**

Byrne, B. M. (1989). *A primer of LISREL: Basic applications and programming for confirmatory factor analytic models*. New York: Springer-Verlag.
Byrne, B. M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming*. Mahwah, NJ: Lawrence Erlbaum.
Diamantopoulos, A., & Siguaw, J. A. (2000). *Introducing LISREL: A guide for the uninitiated*. Thousand Oaks, CA: Sage.
Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Chicago: SSI Scientific.
Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: User's reference guide*. Chicago: SSI Scientific.
Kelloway, K. E. (1998). *Using LISREL for structural equation modeling: A researcher's guide*. Thousand Oaks, CA: Sage.

**Websites**

Scientific Software, International: http://www.ssicentral.com

# Literature Review

Literature reviews are systematic syntheses of previous work around a particular topic. Nearly all scholars have written literature reviews at some point; such reviews are common requirements for class projects or as part of theses, are often the first section of empirical papers, and are sometimes written to summarize a field of study. Given the increasing amount of literature in many fields, reviews are critical in synthesizing scientific knowledge. Although common and important to science, literature reviews are rarely considered to be held to the same scientific rigor as other aspects of the research process. This entry describes the types of literature reviews and scientific standards for conducting literature reviews.

## Types of Literature Reviews

Although beginning scholars often believe that there is one predefined approach, various types of literature reviews exist. Literature reviews can vary along at least seven dimensions.

### *Focus*

The focus is the basic unit of information that the reviewer extracts from the literature. Reviews most commonly focus on research outcomes, drawing conclusions of the form of "The research shows X" or "These studies find X whereas other studies find Y." Although research outcomes are most common, other foci are possible. Some reviews focus on research methods, for example, considering how many studies in a field use longitudinal designs. Literature reviews can also focus on theories, such as what theoretical explanations are commonly used within a field or attempts to integrate multiple theoretical perspectives. Finally, literature reviews can focus on typical practices within a field, for instance, on what sort of interventions are used in clinical literature or on the type of data analyses conducted within an area of empirical research.

### *Goals*

Common goals include integrating literature by drawing generalizations (e.g., concluding the strength of an effect from several studies), resolving conflicts (e.g., why an effect is found in some studies but not others), or drawing links across separate fields (e.g., demonstrating that two lines of research are investigating a common phenomenon). Another goal of a literature review might be to identify central issues, such as unresolved questions or next steps for future research. Finally, some reviews have the goal of criticism; although this goal might sound unsavory, it is important for scientific fields to be evaluated critically and have shortcomings noted.

### *Perspective*

Literature reviews also vary in terms of perspective, with some attempting to represent the literature neutrally and others arguing for a position. Although few reviews fall entirely on one end of this dimension or the other, it is useful for readers to consider this perspective when evaluating a review and for writers to consider their own perspective.

### *Coverage*

Coverage refers to the amount of literature on which the review is based. At one extreme of this dimension is exhaustive coverage, which uses all available literature. A similar approach is the exhaustive review with selective citation, in which the reviewer uses all available literature to draw conclusions but cites only a sample of this literature when writing the review. Moving along this dimension, a review can be representative, such that the reviewer bases conclusions on and cites a subset of the existing literature believed to be similar to the larger body of work. Finally, at the far end of this continuum is the literature review of most central works.

### *Organization*

The most common organization is conceptual, in which the reviewer organizes literature around specific sets of findings or questions. However, historic organizations are also useful, in that they provide a perspective on how knowledge or practices have changes across time. Methodological organizations, in which findings are arranged according to methodological aspects of the reviewed studies, are also a possible method of organizing literature reviews.

### *Method of Synthesis*

Literature reviews also vary in terms of how conclusions are drawn, with the endpoints of this continuum being qualitative versus quantitative. Qualitative reviews, which are also called narrative reviews, are those in which reviewers draw conclusions based on their subjective evaluation of the literature. Vote counting methods, which might be considered intermediate on the qualitative versus quantitative dimension, involve tallying the number of studies that find a particular effect and basing conclusions on this tally. Quantitative reviews, which are sometimes also called meta-analyses, involve assigning numbers to the results of studies (representing an effect size) and then performing statistical analyses of these results to draw conclusions.

### *Audience*

Literature reviews written to support an empirical study are often read by specialized scholars in one's own field. In contrast, many stand-alone reviews are read by those outside one's own field, so it is important that these are accessible to scholars from other fields. Reviews can also serve as a valuable resource for practitioners in one's field (e.g., psychotherapists and teachers) as well as policy makers and the general public, so it is useful if reviews are written in a manner accessible to educated laypersons. In short, the reviewer must consider the likely audiences of the review and adjust the level of specificity and technical detail accordingly.

All of these seven dimensions are important considerations when preparing a literature review. As might be expected, many reviews will have multiple levels of these dimensions (e.g., multiple goals directed toward multiple audiences). Tendencies exist for co-occurrence among dimensions; for example, quantitative reviews typically focus on research outcomes, cover the literature exhaustively, and are directed toward specialized scholars. At the same time, consideration of these dimensions suggests the wide range of possibilities available in preparing literature reviews.

## Scientific Standards for Literature Reviews

Given the importance of literature reviews, it is important to follow scientific standards in preparing these reviews. Just as empirical research follows certain practices to ensure validity, we can consider how various decisions impact the quality of conclusions drawn in a literature review. This section follows Harris Cooper's organization by describing considerations at five stages of the literature review process.

### *Problem Formulation*

As in any scientific endeavor, the first stage of a literature review is to formulate a problem. Here, the central considerations involve the questions that the reviewer wishes to answer, the constructs of interest, and the population about which conclusions are drawn. A literature review can only answer questions about which prior work exists.

For instance, to make conclusions of causality, the reviewer will need to rely on experimental (or perhaps longitudinal) studies; concurrent naturalistic studies would not provide answers to this question. Defining the constructs of interest poses two potential complications: The existing literature might use different terms for the same construct, or the existing literature might use similar terms to describe different constructs. The reviewer, therefore, needs to define clearly the constructs of interest when planning the review. Similarly, the reviewer must consider which samples will be included in the literature review, for instance, deciding whether studies of unique populations (e.g., prison, psychiatric settings) should be included within the review. The advantages of a broad approach (in terms of constructs and samples) are that the conclusions of the review will be more generalizable and might allow for the identification of important differences among studies, but the advantages of a narrow approach are that the literature will likely be more consistent and the quantity of literature that must be reviewed is smaller.

### *Literature Retrieval*

When obtaining literature relevant for the review, it is useful to conceptualize the literature included as a sample drawn from a population of all possible works. This conceptualization highlights the importance of obtaining an unbiased sample of literature for the review. If the literature reviewed is not exhaustive, or at least representative, of the extant research, then the conclusions drawn might be biased. One common threat to all literature reviews is publication bias, or the file drawer problem. This threat is that studies that fail to find significant effects (or that find counterintuitive effects) are less likely to be published and, therefore, are less likely to be included in the review. Reviewers should attempt to obtain unpublished studies, which will either counter this threat or at least allow the reviewer to evaluate the magnitude of this bias (e.g., comparing effects from published vs. unpublished studies). Another threat is that reviewers typically must rely on literature written in a language they know (e.g., English); this excludes literature written in other languages and therefore might exclude most studies conducted in other countries. Although it would be impractical for the reviewer to learn every language in which relevant literature might be written, the reviewer should be aware of this limitation and how it impacts the literature on which the review is based. To ensure transparency of a literature review, the reviewer should report means by which potentially relevant literature was searched and obtained.

### *Inclusion Criteria*

Deciding which works should inform the review involves reading the literature obtained and drawing conclusions regarding relevance. Obvious reasons to exclude works include the investigation of constructs or samples that are irrelevant to the review (e.g., studies involving animals when one is interested in human behavior) or that do not provide information relevant to the review (e.g., treating the construct of interest only as a covariate). Less obvious decisions need to be made with works that involve questionable quality or methodological features different from other studies. Including such works might improve the generalizability of the review on the one hand, but it might contaminate the literature basis or distract focus on the other hand. Decisions at this stage will typically involve refining the problem formulation stage of the review.

### *Interpretation*

The most time-consuming and difficult stage is analyzing and interpreting the literature. As mentioned, several approaches to drawing conclusions exist. Qualitative approaches involve the reviewer performing some form of internal synthesis; as such, they are prone to reviewer subjectivity. At the same time, qualitative approaches are the only option when reviewing nonempirical literature (e.g., theoretical propositions), and the simplicity of qualitative decision making is adequate for many purposes. A more rigorous approach is the vote-counting methods, in which the reviewer tallies studies into different categories (e.g., significant versus nonsignificant results) and bases decisions on either the preponderance of evidence (informal vote counting) or statistical procedures

(comparing the number of studies finding significant results with that expected by chance). Although vote-counting methods reduce subjectivity relative to qualitative approaches, they are limited in that the conclusions reached involve only whether there is an effect (rather than the magnitude of the effect). The best way to draw conclusions from empirical literature is through quantitative, or meta-analytic, approaches. Here, the reviewer codes effect sizes for the studies then applies statistical procedures to evaluate the presence, magnitude, and sources of differences of these effects across studies.

### *Presentation*

Although presentation formats are highly disciplinary specific (and therefore, the best way to learn how to present reviews is to read reviews in one's area), a few guidelines are universal. First, the reviewer should be transparent about the review process. Just as empirical works are expected to present sufficient details for replication, a literature review should provide sufficient detail for another scholar to find the same literature, include the same works, and draw the same conclusions. Second, it is critical that the written report answers the original questions that motivated the review or at least describes why such answers cannot be reached and what future work is needed to provide these answers. A third guideline is to avoid study-by-study listing. A good review synthesizes—not merely lists—the literature (it is useful to consider that a phonebook contains a lot of information, but is not very informative, or interesting, to read). Reviewers should avoid "Author A found . . . Author B found . . ." writing. Effective presentation is critical in ensuring that the review has an impact on one's field.

*Noel A. Card*

***See also*** Effect Size, Measures of; File Drawer Problem; Meta-Analysis

#### Further Readings

Bem, D. J. (1995). Writing a review article for Psychological Bulletin. *Psychological Bulletin, 118,* 172–177.

Card, N. A. (in press). *Meta-analysis: Quantitative synthesis of social science research*. New York: Guilford.

Cooper, H. (1998). *Synthesizing research: A guide for literature reviews* (3rd ed.). Thousand Oaks, CA: Sage.

Cooper, H., & Hedges, L. V. (Eds.). (1994). *The handbook of research synthesis*. New York: Russell Sage Foundation.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.

Pan, M. L. (2008). *Preparing literature reviews: Qualitative and quantitative approaches* (3rd ed.). Glendale, CA: Pyrczak Publishing.

Rosenthal, R. (1995). Writing meta-analytic reviews. *Psychological Bulletin, 118,* 183–192.

## *Logic of Scientific Discovery, The*

*The Logic of Scientific Discovery* first presented Karl Popper's main ideas on methodology, including falsifiability as a criterion for science and the representation of scientific theories as logical systems from which other results followed by pure deduction. Both ideas are qualified and extended in later works by Popper and his follower Imré Lakatos.

Popper was born in Vienna, Austria, in 1902. During the 1920s, he was an early and enthusiastic participant in the philosophical movement called the Vienna Circle. After the rise of Nazism, he fled Austria for New Zealand, where he spent World War II. In 1949, he was appointed Professor of Logic and Scientific Method at the London School of Economics (LSE), where he remained for the rest of his teaching career. He was knighted by Queen Elizabeth II in 1965. Although he retired in 1969, he continued a prodigious output of philosophical work until his death in 1994. He was succeeded at LSE by his protégé Lakatos, who extended his methodological work in important ways.

*The Logic of Scientific Discovery*'s central methodological idea is *falsifiability.* The Vienna Circle philosophers, or logical positivists, had proposed, first, that all meaningful discourse was completely verifiable, and second, that science was coextensive with meaningful discourse. Originally,

they meant by this that a statement should be considered meaningful, and hence scientific, if and only if it was possible to show that it was true, either by logical means or on the basis of the evidence of the senses. Popper became the most important critic of their early work. He pointed out that scientific laws, which are represented as unrestricted or universal generalizations such as "all planets have elliptical orbits" (Kepler's Second Law), are not verifiable by any finite set of sense observations and thus cannot be counted as meaningful or scientific. To escape this paradox, Popper substituted falsifiability for verifiability as the key logical relation of scientific statements. He thereby separated the question of meaning from the question of whether a claim was scientific. A statement could be considered scientific if it could, in principle, be shown to be false on the basis of sensory evidence, which in practice meant experiment or observation. "All planets have elliptical orbits" could be shown to be false by finding a planet with an orbit that was not an ellipse. This has never happened, but if it did the law would be counted as false, and such a discovery might be made tomorrow. The law is scientific because it is falsifiable, although it has not actually been falsified. Falsifiability requires only that the conditions under which a statement would be deemed false are specifiable; it does not require that they have actually come about. However, when this happens, Popper assumed scientists would respond with a new and better conjecture. Scientific methodology should not attempt to avoid mistakes, but rather, as Popper famously put it, it should try to make its mistakes as quickly as possible. Scientific progress results from this sequence of conjectures and refutations, with each new conjecture requiring the precise grounds of specification for its failure to satisfy the principle of falsifiability. Popper's image of science achieved great popularity among working scientists, and he was acknowledged by several Nobel prize winners (including Peter Medewar, John Eccles, and Jacques Monod).

In *The Logic of Scientific Discovery,* Popper, like the logical positivists, presented the view that scientific theories ideally took the form of logically independent and consistent systems of axioms from which (with the addition of initial conditions) all other scientific statements followed by logical deduction. However, in an important later paper ("The Aim of Science," reprinted in *Objective Knowledge,* chapter 5) Popper pointed out that there was no deductive link between, for example, Newton's laws and the original statements of Kepler's laws of planetary motion or Galileo's law of fall. The simple logical model of science offered by Popper and later logical positivists (e.g., Carl Hempel and Paul Oppenheim) therefore failed for some of the most important intertheoretical relations in the history of science.

An additional limitation of falsifiability as presented in *The Logic of Scientific Discovery* was the issue of ad hoc hypotheses. Suppose, as actually happened between 1821 and 1846, a planet is observed that seems to have an orbit that is not an ellipse. The response of scientists at the time was not to falsify Kepler's law, or the Newtonian Laws of Motion and Universal Gravitation from which it was derived. Instead, they deployed a variety of auxiliary hypotheses ad hoc, which had the effect of explaining away the discrepancy between Newton's laws and the observations of Uranus, which led to the discovery of the planet Neptune. Cases like this suggested that any claim could be permanently insulated from falsifying evidence by introducing an ad hoc hypothesis every time negative evidence appeared. Indeed, this could even be done if negative evidence appeared against the ad hoc hypothesis itself; another ad hoc hypothesis could be introduced to explain the failure and so on ad infinitum. Arguments of this type raised the possibility that falsifiability might be an unattainable goal, just as verifiability had been for the logical positivists.

Two general responses to these difficulties appeared. In 1962, Thomas Kuhn argued in *The Structure of Scientific Revolutions* that falsification occurred only during periods of cumulative normal science, whereas the more important noncumulative changes, or revolutions, depended on factors that went beyond failures of observation or experiment. Kuhn made extensive use of historical evidence in his arguments. In reply, Lakatos shifted the unit of appraisal in scientific methodology, from individual statements of law or theory to a historical sequence of successive theories called a research program. Such programs were to be appraised according to whether new additions, ad hoc or otherwise, increased the overall explanatory scope of the program (and especially covered previously unexplained facts) while retaining the

successful content of earlier theories. In addition to *The Logic of Scientific Discovery,* Popper's main ideas are presented in the essays collected in *Conjectures and Refutations* and *Objective Knowledge*. Two books on political philosophy, *The Open Society and Its Enemies* and *The Poverty of Historicism* were also important in establishing his reputation. A three-volume *Postscript to the Logic of Scientific Discovery,* covering respectively, realism, indeterminism, and quantum theory, appeared in 1982.

*Peter Barker*

***See also*** Hypothesis; Scientific Method; Significance Level, Concept of

**Further Readings**

Hempel, C. G., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science, 15,* 135–175.

Kuhn, T. S. (1962). *The structure of scientific revolutions.* Chicago: Chicago University Press.

Lakatos, I. (1978). *The methodology of scientific research programmes.* Cambridge, UK: Cambridge University Press.

Popper, K. R. (1945). *The open society and its enemies.* London: George Routledge and Sons.

Popper, K. R. (1957). *The poverty of historicism.* London: Routledge and Kegan Paul.

Popper, K. R. (1959). *The logic of scientific discovery* (Rev. ed.). New York: Basic Books.

Popper, K. R. (1972). *Objective knowledge: An evolutionary approach.* Oxford, UK: Clarendon Press.

Popper, K. R. (1982). *Postscript to the logic of scientific discovery* (W. W. Bartley III, Ed.). Totowa, NJ: Rowman and Littlefield.

# Logistic Regression

Logistic regression is a statistical technique used in research designs that call for analyzing the relationship of an outcome or *dependent variable* to one or more predictors or *independent variables* when the dependent variable is either (a) *dichotomous,* having only two categories, for example, whether one uses illicit drugs (no or yes); (b) *unordered polytomous,* which is a nominal scale variable with three or more categories, for example, political party identification (Democrat, Republican, other, or none); or (c) *ordered polytomous,* which is an ordinal scale variable with three or more categories, for example, level of education completed (e.g., less than elementary school, elementary school, high school, an undergraduate degree, or a graduate degree). Here, the basic logistic regression model for dichotomous outcomes is examined, noting its extension to polytomous outcomes and its conceptual roots in both loglinear analysis and the general linear model. Next, consideration is given to methods for assessing the goodness of fit and predictive utility of the overall model, and calculation and interpretation of logistic regression coefficients and associated inferential statistics to evaluate the importance of individual predictors in the model. The discussion throughout the entry assumes an interest in prediction, regardless of whether causality is implied; hence, the language of "outcomes" and "predictors" is preferred to the language of "dependent" and "independent" variables.

The equation for the logistic regression model with a dichotomous outcome is

$$\text{logit}(Y) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K,$$

where $Y$ is the dichotomous outcome; logit($Y$) is the natural logarithm of the odds of $Y$, a transformation of $Y$ to be discussed in more detail momentarily; and there are $k = 1, 2, \ldots, K$ predictors $X_k$ with associated coefficients $\beta_k$, plus a constant or *intercept* $\alpha$, which represents the value of logit($Y$) when all of the $X_k$ are equal to zero. If the two categories of the outcome are coded 1 and 0, respectively, and $P_1$ is the probability of being in the category coded as 1, and $P_0$ is the probability of being in the category coded as 0, then the odds of being in category 1 are

$$P_1/P_0 = P_1/(1 - P_1)$$

(because the probability of being in one category is one minus the probability of being in the other category). Logit($Y$) is the natural logarithm of the odds,

$$\ln[P_1/(1 - P_1)],$$

where ln represents the natural logarithm transformation.

## Polytomous Logistic Regression Models

When the outcome is polytomous, logistic regression can be implemented by splitting the outcome into a set of dichotomous variables. This is done by means of contrasts, which identify a reference category (or set of categories) with which to compare each of the other categories (or sets of categories). For a nominal outcome, the most commonly used model is called the baseline category logit model. In this model, the outcome is divided into a set of dummy variables, each representing one of the categories of the outcome, with one of the categories designated as the reference category, in the same way that dummy coding is used for nominal predictors in linear regression. If there are $M$ categories in the outcome, then

$$\text{logit}(Y_m) = \ln(P_m/P_0) = \alpha_m + \beta_{1,m}X_1 + \beta_{2,m}X_2 + \cdots + \beta_{K,m}X_K,$$

where $P_0$ is the probability of being in the reference category and $P_m$ is the probability of being in category $m = 1, 2, \ldots, M-1$, given that the case is either in category $m$ or in the reference category. A total of $M - 1$ equations or logit functions are thus estimated, each with its own intercept $\alpha_m$ and logistic regression coefficients $\beta_{k,m}$, representing the relationship of the predictors to $\text{logit}(Y_m)$.

For ordinal outcomes, the situation is more complex, and several different contrasts might be used. In the adjacent category logit model, for example, each category is contrasted only with the single category preceding it. In the cumulative logit model, (a) for the first logit function, the first category is contrasted with all of the categories following it, then (b) for the second logit function, the first two categories are contrasted with all of the categories following them, and so forth, until for the last $(M - 1)$ logit function, all the categories preceding the last are contrasted with the last category. Other contrasts are also possible. The cumulative logit model is the model most commonly used in logistic regression analysis for an ordinal outcome, and it has the advantage over other contrasts that splitting or combining categories (representing more precise or cruder ordinal measurement) should not affect estimates for categories other than the categories that are actually split or combined. This property is not characteristic of other ordinal contrasts. It is commonly assumed in ordinal logistic regression that only the intercepts (or thresholds, which are similar to intercepts) differ across the logit functions. The ordinal logistic regression equation can be written (here in the format using intercepts instead of thresholds) as

$$\text{logit}(Y_m) = \alpha_m + \beta_1X_1 + \beta_2X_2 + \cdots + \beta_KX_K,$$

where

$$\alpha_m = \alpha_1, \alpha_2, \ldots, \alpha_{M-1}$$

are the intercepts associated with the $M - 1$ logit functions, but $\beta_1, \beta_2, \ldots, \beta_K$ are assumed to be identical for the $M - 1$ logit functions. This assumption can be tested and, if necessary, modified.

## Logistic Regression, Loglinear Analysis, and the General Linear Model

Logistic regression can be derived from two different sources, the general linear model for linear regression and the logit model in loglinear analysis. Linear regression is used to analyze the relationship of an outcome to one or more predictors when the outcome is a continuous interval or ratio scale variable. Linear regression is used extensively in the analysis of outcomes with a natural metric, such as kilograms, dollars, or numbers of people, where the unit of measurement is such that it makes sense to talk about larger or smaller differences between cases (the difference between the populations of France and Germany is smaller than the difference between the populations of France and China). Usually, it also makes sense to talk about one value being some number of times larger than another ($10,000 is twice as much as $5,000); these comparisons are not applicable to the categorical outcome variables for which logistic regression is used. The equation for linear regression is

$$Y = \alpha + \beta_1X_1 + \beta_2X_2 + \cdots + \beta_KX_K,$$

and the only difference from the logistic regression equation is that the outcome in linear regression is $Y$ instead of logit($Y$). The coefficients $\beta_K$ and intercept $\alpha$ in linear regression are most commonly estimated using ordinary least-squares (OLS) estimation, although other methods of estimation are possible.

For OLS estimation and for statistical inferences about the coefficients, certain assumptions are required, and if the outcome is a dichotomy (or a polytomous variable represented as a set of dichotomies) instead of a continuous interval/ratio variable, several of these assumptions are violated. For a dichotomous outcome, the predicted values might lie outside the range of possible values (suggesting probabilities greater than one or less than zero), especially when there are continuous interval or ratio scale predictors in the model. Inferential statistics are typically incorrect because of heteroscedasticity (unequal residual variances for different values of the predictors) and non-normal distribution of the residuals. It is also assumed that the relationship between the outcome and the predictors is linear; however, in the general linear model, it is often possible to linearize a nonlinear relationship by using an appropriate nonlinear transformation. For example, in research on income (measured in dollars), it is commonplace to use the natural logarithm of income as an outcome, because the relationship of income to its predictors tends to be nonlinear (specifically, logarithmic). In this context, the logit transformation is just one of many possible linearizing transformations.

An alternative to the use of linear regression to analyze dichotomous and polytomous categorical outcomes is logit analysis, which is a special case of loglinear analysis. In loglinear analysis, it is assumed that the variables are categorical and can be represented by a contingency table with as many dimensions as there are variables, with each case located in one cell of the table, corresponding to the combination of values it has on all of the variables. In loglinear analysis, no distinction is made between outcomes and predictors, but in logit analysis, one variable is designated as the outcome, and the other variables are treated as predictors. Each unique combination of values of the predictors represents a covariate pattern. Logit model equations are typically presented in a format different from that used in linear regression and logistic regression, and loglinear and logit models are commonly estimated using iterative maximum likelihood (ML) estimation, in which one begins with a set of initial values for the coefficients in the model, examines the differences between observed and predicted values produced by the model (or some similar criterion), and uses an algorithm to adjust the estimates to improve the model. This process of estimation and adjustment of estimates is repeated in a series of steps (iterations) that end when, to some predetermined degree of precision, there is no change in the fit of the model, the coefficients in the model, or some similar criterion.

Logistic regression can be viewed either as a special case of the general linear model involving the logit transformation of the outcome or as an extension of the logit model to incorporate continuous as well as categorical predictors. The basic form of the logistic regression equation is the same as for the linear regression equation, but the outcome logit($Y$) has the same form as the outcome in logit analysis. The use of the logit transformation ensures that predicted values cannot exceed observed values (for an individual case, the logit of $Y$ is either positive or negative infinity, $+\infty$ or $-\infty$), but it also makes it impossible to estimate the coefficients in the logistic regression equation using OLS. Estimation for logistic regression, as for logit analysis, requires an iterative technique, most often ML, but other possibilities include iteratively reweighted least squares, with roots in the general linear model, or some form of quasi-likelihood or partial likelihood estimation, which might be employed when data are clustered or nonindependent. Common instances of nonindependent data include multilevel analysis, complex sampling designs (e.g., multistage cluster sampling), and designs involving repeated measurement of the same subjects or cases, as in longitudinal research. Conditional logistic regression is a technique for analyzing related samples, for example, in matched case-control studies, in which, with some minor adjustments, the model can be estimated using ML.

## Assumptions of Logistic Regression

Logistic regression assumes that the functional form of the equation is correct, and hence, the

predictors $X_k$ are linearly and additively related to logit($Y$), but variables can be transformed to adjust for nonadditivity and nonlinearity (e.g., nonlinearly transformed predictors or interaction terms). It also assumes that each case is independent of all the other cases in the sample, or when cases are not independent, adjustments can be made in either the estimation procedure or the calculation of standard errors (or both) to adjust for the nonindependence. Like linear regression, logistic regression assumes that the variables are measured without error, that all relevant predictors are included in the analysis (otherwise the logistic regression coefficients might be biased), and that no irrelevant predictors are included in the analysis (otherwise standard errors of the logistic regression coefficients might be inflated). Also as in linear regression, no predictor may be perfectly collinear with one or more of the other predictors in the model. Perfect collinearity means that a predictor is completely determined by or predictable from one or more other predictors, and when perfect collinearity exists, an infinite number of solutions is available that maximize the likelihood in ML estimation or minimize errors of prediction more generally. Logistic regression also assumes that the errors in prediction have a binomial distribution, but when the number of cases is large, the binomial distribution approximates the normal distribution. Various diagnostic statistics have been developed and are readily available in existing software to detect violations of assumptions and other problems (e.g., outliers and influential cases) in logistic regression.

## Goodness of Fit and Accuracy of Prediction

In logistic regression using ML (currently the most commonly used method of estimation), in place of the sum of squares statistics used in linear regression, there are *log likelihood* statistics, which are calculated based on observed and predicted probabilities of being in the respective categories of the outcome variable. When multiplied by $-2$, the difference between two log likelihood statistics has an approximate chi-square distribution for sufficiently large samples involving independent observations. One can construct $-2$ log likelihood statistics (here and elsewhere designated as $D$) for (a) a model with no predictors $D_0$ and (b) the tested model, the model for which the coefficients are actually estimated $D_M$. $D_M$, which is sometimes called the deviance statistic, has been used as a goodness-of-fit statistic, but it has somewhat fallen out of favor because of concerns with alternative possible definitions for the saturated model (depending on whether individual cases or covariate patterns are treated as the units of analysis), and the concern that, for data in which there are few cases per covariate pattern, $D_M$ does not really have a chi-square distribution. The Hosmer–Lemeshow goodness-of-fit index is constructed by grouping the data, typically into deciles, based on predicted values of the outcome. This technique is applicable even with few cases per covariate pattern. There seems to be a trend away from concern with goodness of fit, however, to focus instead on the model chi-square statistic,

$$G_M = D_0 - D_M,$$

which compares the tested model to the model with no predictors. $G_M$ generally does follow a chi-square distribution in large samples and it is analogous to the multivariate $F$ statistic in linear regression and analysis of variance. $G_M$ provides a test of the statistical significance of the overall model in predicting the outcome. An alternative to $G_M$ for models not estimated using ML is the multivariate Wald statistic.

There is a substantial literature on coefficients of determination for logistic regression, in which the goal is to find a measure analogous to $R^2$ in linear regression. When the concern is with how close the predicted probabilities of category membership are to observed category membership (quantitative prediction), two promising options are the likelihood ratio $R^2$ statistic,

$$R_L^2 = G_M/D_0,$$

which is applicable specifically when ML estimation is used, and the OLS $R^2$ statistic itself, which is calculated by squaring the correlation between observed values (coded zero and one) and the predicted probabilities of being in category 1. Advantages of $R_L^2$ include the following: (a) it is based on the quantity actually being maximized in ML estimation, (b) it seems to be uncorrelated with the base rate (the percentage of cases in category 1), and (c) it can be calculated for polytomous as well

as dichotomous outcomes. Other $R^2$ analogs have been proposed but have various problems that include correlation with the base rate (to the extent that the base rate itself seems to determine the calculated accuracy of prediction), having no reasonable value for perfect prediction or for perfectly incorrect prediction, or being limited to dichotomous outcomes.

Alternatively, instead of being concerned with predicted probabilities, one might be concerned with how accurately cases are qualitatively classified into the categories of the outcome by the predictors (qualitative prediction). For this purpose, there is a family of indices of predictive efficiency, which is designated lambda-*p*, tau-*p*, and phi-*p*, that are specifically applicable to qualitative prediction, classification, and selection tables (regardless of whether they were generated by logistic regression or some other technique), as opposed to contingency tables more generally. Finally, none of the aforementioned indices of predictive efficiency (or $R^2$ analogs) takes into account the ordering in an ordered polytomous outcome, for which one would naturally consider ordinal measures of association. Kendall's tau-*b* is an ordinal measure of association that, when squared ($\tau_b{}^2$), has a proportional reduction in error (PRE) interpretation, and it seems most promising for use with ordinal outcomes in logistic regression. Tests of statistical significance can be computed for all these coefficients of determination.

## Unstandardized and Standardized Logistic Regression Coefficients

Interpretation of unstandardized logistic regression coefficients ($b_k$, the estimated value of $\beta_k$) is straightforward and parallel to the interpretation of unstandardized coefficients in linear regression: A one-unit increase in $X_k$ is associated with a $b_k$ increase in logit($Y$) (not in $Y$ itself). If we raise the base of the natural logarithm, $e = 2.718\ldots$, to the power $b_k$, we obtain the odds ratio, here designated $\omega_k$, which is sometimes presented in place of or in addition to $b_k$ and can be interpreted as indicating that a one-unit increase in $X_k$ multiplies the odds of being in category 1 by $\omega_k$. Both $b_k$ and $\omega_k$ convey exactly the same information, just in a different form. There are several possible tests of statistical significance for unstandardized logistic regression coefficients. The univariate Wald statistic can be calculated either as the ratio of the logistic regression coefficient to its standard error (*SE*),

$$b_k/SE(b_k),$$

which has an approximate normal distribution, or $[b_k/SE(b_k)]^2$, which has an approximate chi-square distribution. The Wald statistic, however, tends to be problematic for large $b_k$, tending to fail to reject the null hypothesis when the null hypothesis is false (Type II error), but it might still be the best available option when ML is not used to estimate the model. Alternatives include the score statistic and the likelihood ratio statistic (the latter being the difference in $D_M$ with and without $X_k$ in the equation). When ML estimation is used, the likelihood ratio statistic, which has a chi-square distribution and applies to both $b_k$ and $\omega_k$, is generally the preferred test of statistical significance for $b_k$ and $\omega_k$.

Unless all predictors are measured in exactly the same units, neither $b_k$ nor $\omega_k$ clearly indicates whether one variable has a stronger impact on the outcome than another. Likewise, the statistical significance of $b_k$ or $\omega_k$ tells us only how sure we are that a relationship exists, not how strong the relationship is. In linear regression, to compare the substantive significance (strength of relationship, which does not necessarily correspond to statistical significance) of predictors measured in different units, we often rely on standardized regression coefficients. In logistic regression, there are several alternatives for obtaining something like a standardized coefficient. A relatively quick and easy option is simply to standardize the predictors (standardizing the outcome does not matter, because it is the probability of being in a particular category of $Y$, not the actual value of $Y$, that is predicted in logistic regression). A slightly more complicated approach is to calculate

$$b_k^* = (b_k)(s_x)(R)/s_{\text{logit}(Y)},$$

where $b_k^*$ is the fully standardized logistic regression coefficient, $b_k$ is the unstandardized logistic regression coefficient, $s_x$ is the standard deviation of the predictor $X_k$, $R$ is the correlation between the observed value of $Y$ and the predicted probability of being in category 1 of $Y$, $s_{\text{logit}(Y)}$ is the

standard deviation of the predicted values of logit($Y$), and the quantity $s_{\text{logit}(Y)}/R$ represents the estimated standard deviation in the observed values of logit($Y$) (which must be estimated, because the observed values are positive or negative infinity for any single case). The advantage to this fully standardized logistic regression coefficient is that it behaves more like the standardized coefficient in linear regression, including showing promise for use in path analysis with logistic regression. This technique is currently under development. Also, parallel to the use of OLS regression or more sophisticated structural equation modeling techniques in linear panel analysis, it is possible to use logistic regression in panel analysis; once one decides on an appropriate way to measure change in the linear panel analysis, the application of logistic regression is straightforward.

### Logistic Regression and Its Alternatives

Alternatives to logistic regression include probit analysis, discriminant analysis, and models practically identical to the logistic regression model but with different distributional assumptions (e.g., complementary log-log or extreme value instead of logit). Logistic regression, however, has increasingly become the method most often used in empirical research. Its broad applicability to different types of categorical outcomes and the ease with which it can be implemented in statistical software algorithms, plus its apparent consistency with realistic assumptions about real-world empirical data, have led to the widespread use of logistic regression in the biomedical, behavioral, and social sciences.

*Scott Menard*

***See also*** Chi-Square Test; Coefficients of Correlation, Alienation, and Determination; Collinearity; Dependent Variable; Dummy Coding; ***F*** Test; General Linear Model; Independent Variable; Interaction; Least Squares, Methods of; Likelihood Ratio Statistic; Multiple Regression; Odds Ratio; Significance, Statistical

#### Further Readings

Fienberg, S. E. (1980). *The analysis of cross-classified categorical data* (2nd ed.). Cambridge: MIT Press.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York: John Wiley.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). London: Chapman & Hall.

Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. *The American Statistician, 54,* 17–24.

Menard, S. (2002). *Applied logistic regression analysis* (2nd ed.). Thousand Oaks, CA: Sage.

Menard, S. (2004). Six approaches to calculating standardized logistic regression coefficients. *The American Statistician, 58,* 218–223.

Menard, S. (2008). Panel analysis with logistic regression. In S. Menard (Ed.), *Handbook of longitudinal research: Design, measurement, and analysis.* San Francisco: Academic Press.

O'Connell, A. A. (2006). *Logistic regression models for ordinal response variables.* Thousand Oaks, CA: Sage.

Pregibon, D. (1981). Logistic regression diagnostics. *Annals of Statistics, 9,* 705–724.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Simonoff, J. S. (1998). Logistic regression, categorical predictors, and goodness of fit: It depends on who you ask. *The American Statistician, 52,* 10–14.

## Loglinear Models

This entry provides a nontechnical description of loglinear models, which were developed to analyze multivariate cross-tabulation tables. Although a detailed exposition is beyond its scope, the entry describes when loglinear models are necessary, what these models do, how they are tested, and the more familiar extensions of binomial and multinomial logistic regression.

### Why Loglinear Models?

Many social science phenomena, such as designated college major or type of exercise, are nonnumeric, and categories of the variable cannot even be ordered from highest to lowest. Thus, the phenomenon is a *nominal dependent variable;* its categories form a set of mutually exclusive qualities or traits. Any two cases might fall into the same or different categories, but we cannot assert that the value of one case is more or less than that of a second.

Many popular statistics assume the dependent or criterion variable is numeric (e.g., years of formal education). What can the analyst investigating a nominal dependent variable do? There are several techniques for investigating a nominal dependent variable, many of which are discussed in the next section. (Those described in this entry can also be used with ordinal dependent variables. The categories of an ordinal variable can be rank ordered from highest to lowest, or most to least.)

One alternative is logistic regression. However, many analysts have learned binomial logistic regression using only dichotomous or "dummy" dependent variables scored 1 or 0. Furthermore, the uninitiated interpret logistic regression coefficients as if they were ordinary least squares (OLS) regression coefficients. A second analytic possibility uses three-way cross-tabulation tables and control variables with nonparametric statistical measures. This venerable tradition of "physical" (rather than "statistical") control presents its own problems, as follows:

- Limited inference tests for potential three-variable statistical interactions.
- Limiting the analysis to an independent, dependent, and control variable.
- There is no "system" to test whether one variable affects a second indirectly through a third variable; for example, education usually influences income indirectly through its effects on occupational level.
- The three-variable model has limited utility for researchers who want to compare several causes of a phenomenon.

A third option is the linear probability model (LPM) for a dependent dummy variable scored 1 or 0. In this straightforward, typical OLS regression model, B coefficients are interpreted as raising or lowering the probability of a score of 1 on the dependent variable.

However, the LPM, too, has several problems. The regression often suffers from heteroscedasticity in which the dependent variable variance depends on scores of the independent variable(s). The dependent variable variance is truncated (at a maximum 0.25.) The LPM can predict impossible values for the dependent variable that are larger than 1 or less than 0.

Thus the following dilemma: Many variables researchers would like to explain are non-numeric. Using OLS statistics to analyze them can produce nonsensical or misleading results. Some common methods taught in early statistics classes (e.g., three-way cross tabulations) are overly restrictive or lack tests of statistical significance. Other techniques (e.g., LPM) have many unsatisfactory outcomes.

Loglinear models were developed to address these issues. Although these models have a relatively long history in statistical theory, their practical application awaited the use of high-speed computers.

## What Is a Loglinear Model?

Technically, a loglinear model is a set of specified parameters that generates a multivariate cross-tabulation table of expected frequencies or table cell counts. In the general cell frequency (GCF) loglinear model, interest centers on the joint and simultaneous distribution of several variables in the table cells. The focus includes relationships among independent variables as well as those between an independent and a dependent variable.

Table 1 is a simple four-cell ($2 \times 2$) table using 2008 General Social Survey data (NORC at the University of Chicago), which is an in-person representative sample of the United States. Table 1 compares 1,409 male and female adults on the percentage who did or did not complete a high-school chemistry course.

Although men reported completing high-school chemistry more than women by 8%, these results could reflect sampling error (i.e., they are a "sample accident" not a "real" population sex difference).

**Table 1** Respondent Completed High-School Chemistry Course by Sex

| *Completed High-School Chemistry Course* | *Sex* | |
|---|---|---|
| | *Male* | *Female* |
| Yes | 55.5% | 47.5% |
| No | 45.5 | 52.5 |
| Total | 100.0% (869) | 100.0% (540) |

The loglinear analyst compares the set of generated or expected table cell frequencies with the set of observed table cell counts. If the two sets of cell counts coincide overall within sampling error, the analyst says, "the model fits." If the deviations between the two sets exceed sampling error, the model is a "poor fit." Under the latter circumstances, the analyst must respecify the parameters to generate new expected frequencies that more closely resemble the observed table cell counts.

Even in a two-variable table, more than one outcome model is possible. One model, for example, could specify that in the American population, males and females completed a chemistry course at equal rates; thus, in this sample, we would predict that 52.5% of each sex completed high-school chemistry. This outcome is less complicated than one specifying sex differences: If females and similarly completed high-school chemistry, then explaining a sex difference in chemistry exposure is unnecessary.

Table 2 shows expected frequency counts for this "no sex differences" model for each cell above the diagonal (with the actual observed frequencies in bold below it). Thus, when calculating expected frequencies, sample males and females were assumed to have 52.5% high-school chemistry completion rates. The table has been constrained to match the overall observed frequencies for gender and chemistry course exposure.

Comparing expected and observed cell counts, males have fewer expected than observed cases completing chemistry, whereas females have greater expected than observed cases completing high-school chemistry.

Statistically significant GCF coefficients increase or decrease the predicted (modeled) cell counts in a multivariate cross-tabulation table. Negative parameters mean fewer cell frequencies than would occur with a predicted no-effects model. Positive parameters mean higher cell counts than a no-effects model would predict.

**Table 2** Expected and Observed Frequencies for High-School Chemistry Course by Sex

| *Completed High-School Chemistry Course* | *Sex* | | |
|---|---|---|---|
| | *Male* | *Female* | *Total* |
| Yes | 456/ **483** | 284/ **257** | **740** |
| No | 413/ **386** | 256/ **283** | **669** |
| Total | **869** | **540** | **1409** |

Parameters in loglinear models (and by extension their cousins, logistic regression and logit models) are maximum likelihood estimators (MLEs). Unlike direct estimates such as OLS coefficients in linear regression, MLEs are solved through iterative, indirect methods. Reestimating MLEs, which can take several successively closer reestimate cycles, is why high-speed computers are needed.

## A Basic Building Block of Loglinear Models: The Odds Ratio

The odds ratio is formed by the ratio of one cell count in a variable category to a second cell count for the same variable, for example, the U.S. ratio of males to females. Compared with the focus on the entire table in GCF models, this odds ratios subset of loglinear models focuses on categories of the dependent variable (categories in the entire table, which the loglinear model examines, are used to calculate the odds ratios, but the emphasis is on the dependent variable and less on the table as a whole). In Table 2, 740 adults completed a chemistry course and 669 did not, making the odds ratio or odds yes:no 740/669 or 1.11. An odds ratio of 1 would signify a 50–50 split on completing high-school chemistry for the entire sample.

In a binary odds, one category is designated as a "success" ("1"), which forms the odds numerator, and the second as a "failure" ("0"), which forms the ratio denominator. These designations do not signify any emotive meaning of "success." For example, in disease death rates, the researcher might designate death as a success and recovery as a failure. The odds can vary from zero (no successes) to infinity; they are undefined when the denominator is zero. The odds are fractional when there are more failures than successes; for example, if most people with a disease survive, then the odds would be fractional.

A first-order conditional odds considers one independent variable as well as scores on the

dependent variable. The observed first-order chemistry conditional for males in Table 2 (yes:no) is 483/386 = 1.25, and for females it is 257/283 = 0.91. Here, the first-order conditional indicates that males more often completed chemistry than not; however, women completed chemistry less often than successfully completed it.

A second-order odds of 1 designates statistical independence; that is, changes in the distribution of the second variable are not influenced by any systematic change in the distribution of the first variable. Second-order odds ratios departing from 1 indicate two variables are associated. Here, the second-order odds (males:females) of the two first-order conditionals on the chemistry course is 1.25/0.91 or 1.37. Second-order odds greater than 1 indicate males completed a chemistry course more often than females, whereas fractional odds would signify women more often completed chemistry. By extension with more variables, third, fourth, or higher order odds ratios can be calculated.

The natural logarithm (base e, or Euler's constant, abbreviated as ln) of the odds is a logit. The male first-order logit on completing chemistry is ln 1.25 or 0.223; for females, it is ln 0.91 = −.094. Positive logits signify more "successes" than "failures," whereas negative logits indicate mostly failures. Unlike the odds ratio, logits are symmetric around zero. An overwhelming number of "failures" would produce a large negative logit. Logits of 0 indicate statistical independence.

Logits can be calculated on observed or modeled cell counts. Analysts more often work with logits when they have designated a dependent variable. Original model effects, including logits, are multiplicative and nonlinear. Because these measures were transformed through logarithms, they become additive and linear, hence the term *loglinear*.

Loglinear parameters for the cross-tabulation table can specify univariate distributions and two variable or higher associations. In the Table 2 independence model, parameters match the observed total case base ($n$) and both univariate distributions exactly. The first-order odds for females and males are set to be identical, forcing identical percentages on the chemistry question (here 52.5%) for both sexes. The second-order odds (i.e., the odds of the first-order odds) are set to 1 and its ln to zero—signifying no sex effect on high-school chemistry completion.

## Testing Loglinear Models

Although several models might be possible in the same table of observed data, not all models will replicate accurately the observed table cell counts within sampling error. A simpler model that fits the data well (e.g., equal proportions of females and males completed high-school chemistry) is usually preferred to one more complex (e.g., males more often elect chemistry than females). Loglinear and logit models can have any number of independent variables; the interrelationships among those and with a dependent variable can quickly become elaborate. Statistical tests estimate how closely the modeled and observed data coincide.

Loglinear and logit models are tested for statistical significance with a likelihood ratio chi-square statistic, sometimes designated $G^2$ or $L^2$, distinguishing it from the familiar Pearson chi-square ($\chi^2$). This multivariate test of statistical significance is one feature that turns loglinear analysis into a system, which is comparable with an $N$-way analysis of variance or multiple regression as opposed to physical control and inspecting separate partial cross tabulations. One advantage of the logarithmic $L^2$ statistic is that it is additive: The $L^2$ can be partitioned with portions of it allocated to different pieces of a particular model to compare simpler with more complicated models on the same cross-tabulation table.

Large $L^2$s imply sizable deviations between the modeled and observed data, which means the loglinear model does not fit the observed cell counts. The analyst then adds parameters (e.g., a sex difference on high school chemistry) to the loglinear equation to make the modeled and observed cell frequencies more closely resemble each other. The most complex model, the fully saturated model, generates expected frequencies that exactly match the observed cell frequencies (irrespective of the number of variables analyzed). The saturated model always fits perfectly with an $L^2 = 0$.

The analyst can test whether a specific parameter or effect (e.g., a sex difference on high-school chemistry) must be retained so the model fits or

whether it can be dropped. The parameter of interest is dropped from the equation; model cell counts are reestimated and the model is retested. If the resulting $L^2$ is large, the respecified model is a poor fit and the effect is returned to the loglinear equation. If the model with fewer effects fits, the analyst next examines which additional parameters can be dropped. In addition to the $L^2$, programs such as SPSS, an IBM product (formerly called PASW Statistics), report *z* scores for each specified parameter to indicate which parameters are probably necessary for the model.

Most models based on observed data are hierarchical; that is, more complex terms contain all lower order terms. For example, in the sex-chemistry four-cell table, a model containing a sex by chemistry association would also match the sex distribution, the chemistry distribution (matching the modeled univariate distribution on the chemistry course to the observed split in the variable), and the case base *n* to the observed sample size. Nonhierarchical models can result from some experimental designs (equal cases for each treatment group) or disproportionate sampling designs. Final models are described through two alternative terminologies. The saturated hierarchical model for Tables 1 and 2 could be designated as (A*B) or as {AB}. Either way, this hierarchical model would include parameters for *n*, the A variable, the B variable, and the AB association. For hierarchical models, lower order terms are assumed included in the more complex terms. For nonhierarchical models, the analyst must separately specify all required lower order terms.

## Degrees of Freedom

$L^2$ statistics are evaluated for statistical significance with respect to their associated degrees of freedom (*df*). The *df* in loglinear models depend on the number of variables, the number of categories in each variable, and the effects the model specifies.

The total *df* depends on the total number of cells in the table. In the saturated $2 \times 2$ (four-cell) table depicted in Tables 1 and 2, each variable has two categories. The case base (*n*) counts as 1 *df*, the variables sex and the chemistry course each have $2-1$ *df*, and the association between sex and the chemistry course has $(2-1)*(2-1)$ or 1 *df*.

Any time we "fix" a parameter, that is, specify that the expected and observed cell counts or variable totals must match for that variable or association, we lose *df*. A fully saturated model specifying a perfect match for all cells has zero *df*. The model fits but might be more complex than we would like.

## Extensions and Uses of Loglinear Models

Logit and logistic regression models are derived from combinations of cells from an underlying GCF model. When the equations for cell counts are converted to odds ratios, terms describing the distributions of and associations among the independent variables cancel and drop from the equation, leaving only the split on the dependent variable and the effects of independent variables on the dependent variable. Because any variable, including a dependent variable, in a GCF model can have several categories, the dependent variable in logistic regression can also have several categories. This is multinomial logistic regression and it extends the more familiar binomial logistic regression.

Of the possible loglinear, logit, and logistic regression models, the GCF model allows the most flexibility, despite its more cumbersome equations. Associations among all variables, including independent variables, can easily be assessed. The analyst can test path-like causal models and check for indirect causal effects (mediators) and statistical interactions (moderators) more readily than in extensions of the GCF model, such as logit models.

Although the terminology and underlying premises of the loglinear model might be unfamiliar to many analysts, it provides useful ways of analyzing nominal dependent variables that could not be done otherwise. Understanding loglinear models also helps to describe correctly the logarithmic (logit) or multiplicative and exponentiated (odds ratios) extensions in logistic regression, giving analysts a systemic set of tools to understand the relationships among non-numeric variables.

*Susan Carol Losh*

***See also*** Likelihood Ratio Statistic; Logistic Regression; Nonparametric Statistics; Odds Ratio

**Further Readings**

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). Hoboken, NJ: John Wiley.
Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed.). Hoboken, NJ: John Wiley.
Aldrich, J. H., & Nelson, F. D. (1995). *Linear probability, logit and probit models*. Thousand Oaks, CA: Sage.
DeMaris, A. (1992). *Logit modeling: Practical applications*. Newbury Park, CA: Sage.
Gilbert, N. (1993). *Analyzing tabular data: Loglinear and logistic models*. London: UCL Press.
Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., & Shavelson, R. J. (2007). *Estimating causal effects using experimental and observational designs*. Washington, DC: American Educational Research Association.

# Longitudinal Design

A longitudinal design is one that measures the characteristics of the same individuals on at least two, but ideally more, occasions over time. Its purpose is to address directly the study of individual change and variation. Longitudinal studies are expensive in terms of both time and money, but they provide many significant advantages relative to cross-sectional studies. Indeed, longitudinal studies are essential for understanding developmental and aging-related changes because they permit the direct assessment of within-person change over time and provide a basis for evaluating individual differences in level as separate from the rate and pattern of change as well as the treatment of selection effects related to attrition and population mortality.

## Traditional Longitudinal Designs

Longitudinal designs can be categorized in several ways but are defined primarily on differences in initial sample (e.g., age homogeneous or age heterogeneous), number of occasions (e.g., semiannual or intensive), spacing between assessments (e.g., widely spaced panel designs or intensive measurement designs), and whether new samples are obtained at subsequent measurement occasions (e.g., sequential designs). These design features can be brought together in novel ways to create study designs that are more appropriate to the measurement and modeling of different outcomes, life periods, and in capturing intraindividual variation, change, and events producing such changes.

In contrast to a cross-sectional design, which allows comparisons across individuals differing in age (i.e., birth cohort), a longitudinal design aims to collect information that allows comparisons across time in the same individual or group of individuals. One dimension on which traditional longitudinal designs can be distinguished is their sampling method. In following a group of individuals over time, one might choose to study a particular birth cohort, so that all the research subjects share a single age and historical context. As an extension of this, a variety of sequential designs exists, in which multiple cohorts are systematically sampled and followed over time. K. Warner Schaie's Seattle Longitudinal Study used such a design, in which new samples of the same cohorts are added at each subsequent observational "wave" of the study. More typical, however, is an age-heterogeneous sample design, which essentially amounts to following an initial cross-sectional sample of individuals varying in age (and therefore birth cohort) over time. Cohort-sequential, multiple cohort, and accelerated longitudinal studies are all examples of mixed longitudinal designs. These designs contain information on both initial between-person age differences and subsequent within-person age changes. An analysis of such designs requires additional care to estimate separately the between-person (i.e., cross-sectional) and within-person (i.e., longitudinal) age-related information. Numerous discussions are available regarding the choice of design and the associated strengths and threats to the validity associated with each.

Another dimension along which traditional longitudinal designs can differ is the interval between waves. Typical longitudinal studies reassess participants at regular intervals, with relatively equal one or several-year gaps, but these might vary from smaller (e.g., half-year) to longer (7-year or decade). Variations in this pattern have been used because of funding cycles; for example, intervals within a single study might range from 2 to 5 or more years, and to creative use of opportunities, such as recontacting in later life a sample of

former participants in a study of child development or of army inductees, who were assessed in childhood or young adulthood. Intensive measurement designs, such as daily diary studies and burst measurement designs, are based on multiple assessments within and across days, permitting the analysis of short-term variation and change.

A relevant design dimension associated with longitudinal studies is breadth of measurement. Given the high financial, time, and effort costs it is not unusual for a longitudinal study to be multidisciplinary, or at least to attempt to address a significant range of topics within a discipline. While some studies have dealt with the cost issue by maintaining a very narrow focus, these easily represent the minority.

## Levels of Analysis in Longitudinal Research

For understanding change processes, longitudinal studies provide many advantages relative to cross-sectional studies. Longitudinal data permit the direct estimation of parameters at multiple levels of analysis, each of which is complementary to understanding population and individual change with age. Whereas cross-sectional analyses permit between-person analysis of individuals varying in age, longitudinal follow-up permits direct evaluation of both between-person differences and within-person change.

Information available in cross-sectional and longitudinal designs can be summarized in terms of seven main levels of analysis and inferential scope (shown in italics in the next section). These levels can be ordered, broadly, in terms of their focus, ranging from the population to the individual. The time sampling generally decreases across levels of analysis, from decades for analysis of historical birth cohort effects to days, minutes, or seconds for assessment of highly variable within-person processes.

These levels of analysis are based on a combination of multiple-cohort, between-person, and within-person designs and analysis approaches and all are represented by recent examples in developmental research. *Between-cohort differences,* which is the broadest level, can be examined to evaluate whether different historical contexts (e.g., indicated by birth cohort) have lasting effects on level and on rate of change in functioning in later life. *Population average trends* describe aggregate population change. Trends can be based on between-person differences in age-heterogeneous studies (although confounded with differences related to birth cohort and sample selection associated with attrition and mortality) or on direct estimates of within-person change in studies with longitudinal follow-up, in which case they can be made conditional on survival. *Between-person age differences* can be analyzed in terms of shared age-related variance in variance decomposition and factor models. This approach to understanding aging, however, confounds individual differences in age-related change with average age differences (i.e., between-person age trends), cohort influences, and mortality selection.

Longitudinal models permit the identification of *individual differences in rates of change* over time, which avoids making assumptions of ergodicity—that age differences between individuals and age changes within individuals are equivalent. In these models, time can be structured in many alternative ways. It can be defined as time since an individual entered the study, time since birth (i.e., chronological age), or time until or since occurrence of a shared event such as retirement or diagnosis of disease. Elaboration of the longitudinal model permits estimation of *association among within-person rates of change in different outcomes,* in other words, using multivariate associations among intercepts and change functions to describe the interdependence of change functions. In shorter term longitudinal designs, researchers have emphasized *within-person variation as an outcome* and have examined whether individuals who display greater variability relative to others exhibit this variation generally across different tasks. *Within-person correlations* (i.e., coupling or dynamic factor analysis) are based on the analysis of residuals (after separating intraindividual means and trends) and provide information regarding the correlation of within-time variation in functioning across variables. Each level of analysis provides complementary information regarding population and individual change, and the inferences and interpretations possible from any single level of analysis have distinct and delimited ramifications for understanding developmental and aging-related change.

## Considerations for the Design of Longitudinal Studies

The levels of analysis described previously correspond roughly to different temporal and historical (i.e., birth cohort) sampling frames and range from very long to potentially very short intervals of assessment. The interpretation, comparison, and generalizability of parameters derived from different temporal samplings must be carefully considered and require different types of designs and measurements. The temporal characteristics of change and variation must be taken into account, as different sampling intervals will generally lead to different results requiring different interpretations for both within and between-person processes. For example, correlations between change and variability over time across outcomes will likely be different for short temporal intervals (minutes, hours, days, or weeks) in contrast to correlations among rates of change across years, the typical intervals of many longitudinal studies on aging.

Measurement interval is also critical for the prediction of outcome variables and for establishing evidence on leading versus lagging indicators. Causal mechanisms need time for their influences to be exerted, and the size of the effect will vary with the time interval between the causal influence and the outcome. Thus, if one statistically controls for a covariate measured at a time before it exerts its causal influence, the resultant model parameters might still be biased by the covariate. Time-varying covariates must be measured within the time frame in which they are exerting their influence to provide adequate representations of the causal, time-dependent processes. However, deciding on what an appropriate time frame might be is not an easy task, and might not be informed by previous longitudinal studies, given that the data collection intervals from many studies are determined by logistical and financial factors, rather than theoretical expectations about the timing of developmental processes.

### *Population Sampling, Attrition, and Mortality*

In observational studies, representative sampling is important, as random assignment to conditions is not possible. However, attrition and mortality selection processes complicate both the definition of an aging population and the sampling procedures relevant to obtaining a representative sample in studies of later life. Attrition in longitudinal studies of aging is often nonrandom, or selective, in that it is likely to result from mortality or declining physical and mental functioning of the participants during the period of observation. This presents an important inferential problem, as the remaining sample becomes less and less representative of the population from which it originated. Generalization from the sample of continuing participants to the initial population might become difficult to justify. However, a major advantage of longitudinal studies is that they contain information necessary to examine the impact of attrition and mortality selection on the observed data. This information, which is inaccessible in cross-sectional data, is essential for valid inferences and improved understanding of developmental and aging processes.

Heterogeneity in terms of chronological age and population mortality poses analytical challenges for both cross-sectional and longitudinal data and is a particular challenge to studies that begin with age-heterogeneous samples. Age-homogeneous studies, where single or narrow age birth cohorts are initially sampled, provide an initially well-defined population that can be followed over time, permitting conditional estimates based on subsequent survival. However, initial sampling of individuals at different ages (i.e., age-heterogeneous samples), particularly in studies of adults and aging, confounds population selection processes related to mortality. The results from longitudinal studies, beginning as age-heterogeneous samples, can be properly evaluated and interpreted when the population parameters are estimated conditional on initial between-person age differences, as well as on mortality and attrition processes that permit inference to defined populations.

Incomplete data can take many forms, such as item or scale nonresponse, participant attrition, and mortality within the population of interest (i.e., lack of initial inclusion or follow-up because of death). Statistical analysis of longitudinal studies is aimed at providing inferences regarding the level and rate of change in functioning, group differences, variability, and construct relations within

a population, and incomplete data complicate this process. To make appropriate population inferences about development and change, it is important not only to consider thoroughly the processes leading to incomplete data (e.g., health, fatigue, cognitive functioning), but also to obtain measurements of these selection and attrition processes to the greatest extent possible and include them in the statistical analysis based on either maximum likelihood estimation or multiple imputation procedures. Longitudinal studies might be hindered by missing data or not being proximal to critical events that represent or influence the process of interest. As a consequence, some researchers have included additional assessments triggered by a particular event or response.

### *Effects of Repeated Testing*

Retest (i.e., practice, exposure, learning, or reactivity) effects have been reported in several longitudinal studies, particularly in studies on aging and cognition where the expected effects are in the opposite direction. Estimates of longitudinal change might be exaggerated or attenuated depending on whether the developmental function is increasing or decreasing with age. Complicating matters is the potential for improvement to occur differentially, related to ability level, age, or task difficulty, as well as to related influences such as warm-up effects, anxiety, and test-specific learning.

Intensive measurement designs, such as those involving measurement bursts with widely spaced sets of intensive measurements, are required to distinguish short-term learning gains from long-term aging-related changes. The typical longitudinal design used to estimate developmental or aging functions usually involves widely spaced intervals between testing occasions. Design characteristics that are particularly sensitive to the assessment of time-related processes, such as retest or learning effects, have been termed *temporal layering* and involve the use of different assessment schedules within longitudinal design (i.e., daily, weekly, monthly, semiannually, or annually). For example, one such alternative, the measurement burst design, where assessment bursts are repeated over longer intervals, is a compromise between single-case time series and conventional longitudinal designs, and they permit the examination of within-person variation, covariation, and change (e.g., because of learning) within measurement bursts and evaluation of change in maximal performance over time across measurement bursts.

### *Selecting Measurement Instruments*

The design of future longitudinal studies on aging can be usefully informed by the analysis and measurement protocol of existing studies. Such studies, completed or ongoing, provide evidence for informing decisions regarding optimal or essential test batteries of health, cognition, personality, and other measures. Incorporating features of measurement used in previous studies, when possible, would permit quantitative anchoring and essential opportunities for cross-cohort and cross-country comparison.

Comparable measures are essential for cross-study comparison, replication, and evaluation of generalizability of research findings. The similarity of a measure can vary at many levels, and within a single nation large operational differences can be found. When considering cross-cultural or cross-national data sets, these differences can be magnified: Regardless of whether the same measure has been used, differences are inevitably introduced because of language, administration, and item relevance. A balance must be found between optimal similarity of administration, similarity of meaning, and significance of meaning—avoiding unreasonable loss of information or lack of depth. These challenges must clearly be addressed in a collaborative endeavor, but in fact they are also critical to general development of the field, for without some means for comparing research products, our findings lack evidence for reproducibility and generalizability.

## Challenges and Strengths

Longitudinal studies are necessary for explanatory theories of development and aging. The evidence obtained thus far from long-term longitudinal and intensive short-term longitudinal studies indicates remarkable within-person variation in many types of processes, even those once considered highly stable (e.g., personality). From both theoretical and empirical perspectives, between-person differences are a complex function of initial individual

differences and intraindividual change. The identification and understanding of the sources of between-person differences and of developmental and aging-related changes requires the direct observation of within-person change available in longitudinal studies.

There are many challenges for the design and analysis of strict within-person studies and large-sample longitudinal studies, and these will differ according to purpose. The challenges of strict within-person studies include limits on inferences given the smaller range of contexts and characteristics available within any single individual. Of course, the study of relatively stable individual characteristics and genetic differences requires between-person comparison approaches. In general, combinations of within-person and between-person population and temporal sampling designs are necessary for comprehensive understanding of within-person processes of aging because people differ in their responsiveness to influences of all types, and the breadth of contextual influences associated with developmental and aging outcomes is unavailable in any single individual. The strength of longitudinal designs is that they permit the simultaneous examination of within-person processes in the context of between-person variability, between-person differences in change, and between-person moderation of within-person processes.

*Scott M. Hofer and Andrea M. Piccinin*

***See also*** Cross-Sectional Design; Population; Sequential Design; Within-Subjects Design

**Further Readings**

Alwin, D. F., Hofer, S. M., & McCammon, R. (2006). Modeling the effects of time: Integrating demographic and developmental perspectives. In R. H. Binstock & L. K. George (Eds.), *Handbook of aging and the social sciences* (6th ed., pp. 20–38). San Diego, CA: Academic Press.

Baltes, P. B., & Nesselroade, J. R. (1979). History and rationale of longitudinal research. In J. R. Nesselroade & P. B. Baltes (Eds.), *Longitudinal research in the study of behavior and development*. New York: Academic Press.

Hofer, S. M., Flaherty, B. P., & Hoffman, L. (2006). Cross-sectional analysis of time-dependent data: Problems of mean-induced association in age-heterogeneous samples and an alternative method based on sequential narrow age-cohorts. *Multivariate Behavioral Research, 41,* 165–187.

Hofer, S. M., & Hoffman, L. (2007). Statistical analysis with incomplete data: A developmental perspective. In T. D. Little, J. A. Bovaird, & N. A. Card (Eds.), *Modeling ecological and contextual effects in longitudinal studies of human development* (pp. 13–32). Mahwah, NJ: Lawrence Erlbaum.

Hofer, S. M., & Piccinin, A. M. (2009). Integrative data analysis through coordination of measurement and analysis protocol across independent longitudinal studies. *Psychological Methods, 14,* 150–164.

Hofer, S. M., & Sliwinski, M. J. (2006). Design and analysis of longitudinal studies of aging. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (6th ed., pp. 15–37). San Diego, CA: Academic Press.

Nesselroade, J. R. (2001). Intraindividual variability in development within and between individuals. *European Psychologist, 6,* 187–193.

Piccinin, A. M., & Hofer, S. M. (2008). Integrative analysis of longitudinal studies on aging: Collaborative research networks, meta-analysis, and optimizing future studies. In S. M. Hofer & D. F. Alwin (Eds.), *Handbook on cognitive aging: Interdisciplinary perspectives* (pp. 446–476). Thousand Oaks, CA: Sage.

Schaie, K. W., & Hofer, S. M. (2001). Longitudinal studies of aging. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (pp. 53–77). San Diego, CA: Academic Press.

# M

## Main Effects

Main effects can be defined as the average differences between one independent variable (or factor) and the other levels of one or more independent variables. In other words, investigators identify main effects, or how one independent variable influences the dependent variable, by ignoring or constraining the other independent variables in a model. For instance, let us say there is a difference between two levels of independent variable A and differences between three levels of independent variable B. Consequently, researchers can study the presence of both factors separately, as in single-factor experiments. Thus, main effects can be determined in either *single-factor experiments* or *factorial design experiments*. In addition, main effects can be interpreted meaningfully only if the interaction effect is absent. This entry focuses on main effects in factorial design, including analysis of the marginal means.

### Main Effects in Factorial Design

Factorial design is applicable whenever researchers wish to examine the influence of a particular factor among two or more factors in their study. This design is a method for controlling various factors of interest in just one experiment rather than repeating the same experiment for each of the factors or independent variables in the study. If there is no significant interaction between the factors and many factors are involved in the study, testing the main effects with a factorial design likely confers efficiency.

Plausibly, in factorial design, each factor may have more than one level. Hence, the significance of the main effect, which is the difference in the *marginal means* of one factor over the levels of other factors, can be examined. For instance, suppose an education researcher is interested in knowing how gender affects the ability of first-year college students to solve algebra problems. The first variable is gender, and the second variable is the level of difficulty of the algebra problems. The second variable has two levels of difficulty: difficult (proof of algebra theorems) and easy (solution of simple multiple-choice questions). In this example, the researcher uses and examines a $2 \times 2$ factorial design. The number "2" represents the number of levels that each factor has. If there are more than two factors, then the factorial design would be adjusted; for instance, the factorial design may look like $3 \times 2 \times 2$ for three factors with 3 levels versus 2 levels and another 2-level factor. Therefore, a total of three main effects would have to be considered in the study.

In the previous example of $2 \times 2$ factorial design, however, both variables are thought to influence the ability of first-year college students to solve algebra problems. Hence, two main effects can be examined: (1) gender effects, while the level of difficulty effects is controlled and (2) level-of-difficulty effects, while gender effects are

**Table 1** Main Effects of Gender and College Year on IQ Test Scores

| | *College Year* | | | | |
|---|---|---|---|---|---|
| *Gender* | *Freshman* | *Sophomore* | *Junior* | *Senior* | *Marginal Means* |
| Male | 100 | 100 | 110 | 120 | 107.5 |
| Female | 100 | 110 | 115 | 125 | 112.5 |
| Marginal means | 100 | 105 | 112.5 | 122.5 | 110 |

controlled. The hypothesis also can be stated in terms of whether first-year male and female college students differ in their ability to solve the more difficult algebra problems. The hypothesis can be answered by examining the simple main effects of gender or the simple main effects of the second variable (level of difficulty).

## Marginal Means

An easy technique for checking the main effect of a factor is to examine the *marginal means,* or the average difference at each level that makes up the factorial design. The differences between levels in a factor could preliminarily affect the dependent variable. The differences in the marginal means also tell researchers how much, on average, one level of the factor differs from the others in affecting the dependent variable. For instance, Table 1 shows the two-level main effect of gender and the four-level main effect of college year on IQ test points. The marginal means from this $2 \times 4$ factorial design show that there might be a main effect of gender, with an average difference of 5 points, on IQ test scores. Also, there might be a main effect of college year, with differences of 5 to 22.5 points in IQ test scores across college years. To determine whether these point differences are greater than what would be expected from chance, the significance of these mains effects needs to be tested.

The test of main effects significance for each factor is the test of between-subject effects provided by the analysis of variance (ANOVA) table found in many statistical software packages, such as the SPSS (an IBM company, formerly called PASW® Statistics), SAS, and MINITAB. The *F* ratio for the two factors—which is empirically computed from the amount of variance in the dependent variable contributed by these two factors—is the ratio of the relative variance of that particular factor to the random error variance. The larger the *F* ratio (i.e., the larger the relative variance), the more likely that the factor significantly affects the dependent variable. To determine whether the *F* ratio is large enough to show that the main effects are significant, the researcher can compare the *F* ratio with critical *F* by using the critical values table provided in many statistics textbooks. The researcher can also compare the *p* value in the ANOVA table with the chosen significance level, say .05. If $p < .05$, then the effect for that factor on the dependent variable is significant. The marginal means can then be interpreted from these results, that is, which group (e.g., male vs. female, freshman vs. senior, or sophomore vs. senior) is significantly higher or lower than the other groups on that factor. It is important to report the *F* and *p* values, followed by the interpretation of the differences in the marginal means of a factor, especially for the significant main effects on the dependent variable.

The analysis of the main effects of a factor on the dependent variable while other factors are controlled is used when a researcher is interested in looking at the pattern of differences between the levels of individual independent variables. The significant main effects give the researcher information about how much one level of a factor could be more or less over the other levels. The significant main effect, however, is less meaningful when the interaction effect is significant, that is, when there is a significant interaction effect between factors A and B. In that case, the researcher should test the *simple main effects* instead of the main effects on the dependent variable.

*Zairul Nor Deana Binti Md Desa*

***See also*** Analysis of Variance (ANOVA); Factorial Design; Interaction; Simple Main Effects

**Further Readings**

Aitken, M. R. F. (2006). *ANOVA for the behavioural sciences researcher.* Mahwah, NJ: Lawrence Erlbaum.

Green, S. B., & Salkind, N. J. (2007). *Using SPSS for Windows and Macintosh: Analyzing and understanding data* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Keppel, G., & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

Myers, R. H., Montgomery, D. C., & Vining, G. G. (2002). *Generalized linear models: With applications in engineering and the sciences.* New York: Wiley.

# Mann–Whitney *U* Test

The Mann–Whitney *U* Test is a popular test for comparing two independent samples. It is a nonparametric test, as the analysis is undertaken on the rank order of the scores and so does not require the assumptions of a parametric test. It was originally proposed by Frank Wilcoxon in 1945 for equal sample sizes, but in 1947 H. B. Mann and D. R. Whitney extended it to unequal sample sizes (and also provided probability values for the distribution of *U*, the test statistic).

When the null hypothesis is true, and the ranks of the two samples are drawn from the same population distribution, one would expect the mean rank for the scores in one sample to be the same as the mean rank for the scores in the other sample. However, if there is an effect of the independent variable on the scores, then one would expect it to influence their rank order, and hence, one would expect the mean ranks to be different for the two samples.

This entry discusses the logic and calculation of the Mann–Whitney *U* test and the probability of *U*.

## The Logic of the Test

The logic of the test can be seen by an example. A group of children sign up for a tennis camp during summer vacation. At the beginning of the course, a tennis expert examines the children on their tennis ability and records each child's performance. The children are then randomly allocated to one of two tennis coaches, either Coach Alba or Coach Bolt. At the end of the course, after 4 weeks of intensive coaching, the tennis expert again examines the children on the test of their tennis ability and records their performance. The amount of improvement in the child's tennis performance is calculated by subtracting their score at the beginning of the course from the one at the end. An interesting question arises: Is it better to be coached by Coach Alba or by Coach Bolt? Given that the children play on the same tennis courts and follow the same course of study, the only difference between the two groups is the coaching. So does one group improve more than the other?

As we are unsure whether the tennis expert's test scores satisfy parametric assumptions, a Mann–Whitney test is undertaken on the improvement scores to test the hypothesis. In this example, Alba coaches six students and Bolt coaches five. In Alba's group, Juan receives an improvement score of 23, Todd gets 15, Maria 42, Charlene 20, Brad 32, and Shannon 28. In Bolt's group, Grace receives an improvement score of 24, Carl gets 38, Kelly 48, Ron 45, and Danny 35. How do we decide whether one of the coaches achieves the better results? First, the results can be seen more clearly if they are put in a table in rank order, that is, listing the scores in order from least improved (at the bottom) to most improved (at the top):

| *Rank* | *Student Name* | *Improvement Score* | *Coach* |
|---|---|---|---|
| 11 | Kelly | 48 | Bolt |
| 10 | Ron | 45 | Bolt |
| 9 | Maria | 42 | Alba |
| 8 | Carl | 38 | Bolt |
| 7 | Danny | 35 | Bolt |
| 6 | Brad | 32 | Alba |
| 5 | Shannon | 28 | Alba |
| 4 | Grace | 24 | Bolt |
| 3 | Juan | 23 | Alba |
| 2 | Charlene | 20 | Alba |
| 1 | Todd | 15 | Alba |

In the final column of the table, it can be seen that five of Alba's students are in the bottom six places.

The calculation of the Mann–Whitney *U* test is undertaken on the ranks, not the original scores. So the key information from the above table, the

students' rank plus their coach, is shown in the following table (where A indicates Coach Alba and B indicates Coach Bolt):

| Rank | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Coach | A | A | A | B | A | A | B | B | A | B | B |

If there really is a difference between the coaches, we would expect the students from one coach to be in the bottom positions in the rank order and the students from the other coach to be in the top positions. However, if there is no difference between the coaches, we would expect the students to be scattered across the ranks. One way to measure this is to take each rank in turn and consider how many results from the other coach are below it. For example, the student at Rank 5 is coached by Coach Alba. There is only one of Coach Bolt's students below Rank 5, so a score of 1 is given to Alba's student at Rank 5. This is done for all Coach Alba's students. Then a total score for Alba's students is produced by adding up these values. This value is called *U*. The same calculation is done for Coach Bolt's students, to produce a second *U* value. The following table shows the results:

| Rank | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | *U* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Coach | A | A | A | B | A | A | B | B | A | B | B | |
| Alba | 0 | 0 | 0 | | 1 | 1 | | | 3 | | | 5 |
| Bolt | | | | 3 | | | 5 | 5 | | 6 | 6 | 25 |

Notice that the *U* score of 5 indicates that most of Coach Alba's students are near the bottom (there are not many of Bolt's students worse than them) and the much larger *U* value of 25 indicates that most of Coach Bolt's students are near the top of the ranks.

However, consider what the *U* scores would be if every one of Alba's students had made up the bottom six ranks. In this case, none of Alba's students would have been above any of Bolt's students in the ranks, and the *U* value would have been 0. The *U* value for Bolt's students would have been 30.

Now consider the alternative situation, when the students from the two coaches are evenly spread across the ranks. Here the *U* values will be very similar. For example, in the following table, the two *U* values are actually the same:

| Rank | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | *U* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Coach | A | B | A | B | A | B | A | B | A | B | A | |
| Alba | 0 | | 1 | | 2 | | 3 | | 4 | | 5 | 15 |
| Bolt | | 1 | | 2 | | 3 | | 4 | | 5 | | 15 |

So large differences in *U* values indicate a possible difference between the coaches, and similar *U* values indicate little difference between the coaches. Thus, we have a statistic for testing the hypothesis. The calculated *U* values of 25 and 5 indicate a possible difference in the coaches.

## The Calculation of the Test

If we refer to Coach Alba's students as Sample 1 and Coach Bolt's students as Sample 2, then $n_1$, the number of scores in Sample 1, is 6 and $n_2$, the number of scores in Sample 2, is 5.

In the worst possible situation (for Coach Alba, that is!) all Coach Alba's students would take up the bottom six places. The sum of their ranks would be $1+2+3+4+5+6=21$. Now how do Alba's students really do? The actual sum of ranks for Alba's group, Sample 1, is $R_1 = 1+2+3+5+6+9=26$. The difference produces the formula for the Mann–Whitney *U* statistic: $U_1$ equals the sum of actual ranks minus the sum of bottom $n_1$ ranks or, expressed as a formula:

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2} = 26 - 21 = 5.$$

This formula gives us the same value of 5 that was calculated by a different method earlier. But the formula provides a simple method of calculation, without having to laboriously inspect the ranks, as above. (Notice also that the mean rank for Alba's group is $\frac{R_1}{n_1} = \frac{26}{6} = 4.33$; below 6, the middle rank for 11 results. This also provides an indication that Alba's students improve less than Bolt's.)

The *U* statistic can be calculated for Coach Bolt. If Bolt's students had been in the bottom five places, their ranks would have added up to 15 $(1+2+3+4+5)$. In actual fact, the sum of the

ranks of Bolt's students is $4 + 7 + 8 + 10 + 11 = 40$. So $U_2$ equals the sum of actual ranks minus the sum of bottom $n_2$ ranks or, expressed as a formula:

$$U_2 = \mathrm{R}_2 - \frac{n_2(n_2+1)}{2} = 40 - 15 = 25.$$

This is a relatively large value, so Bolt's students are generally near the top of the ranks. (Notice also that the mean rank for Bolt's group is $\frac{\mathrm{R}_2}{n_2} = \frac{40}{5} = 8$; above 6, the middle for 11 ranks.) And the two values of $U$ are $U_1 = 5$ and $U_2 = 25$, the same as produced by the different method earlier.

The sum of two $U$ values will always be $n_1 n_2$, which in this case is 30. While the two $U$ values are quite different from each other, indicating a separation of the samples into the lower and upper ranks, in statistical tests a result is significant only if the probability is less than or equal to the significance level (usually, $p = .05$).

## The Probability of *U*

In a two-tailed prediction, researchers always test the smaller value of $U$ because the probability is based on this value. In the example, there are 11 scores, 6 in one sample and 5 in the other. What is the probability of a $U$ value of 5 or smaller when the null hypothesis is true? There are 462 possible permutations for the rank order of 6 scores in one sample and 5 in the other ($\frac{11!}{5!6!}$). Only two permutations produce a $U$ value of zero: all one sample in the bottom ranks or all the other sample in the bottom ranks. It is not difficult to work out (using a mathematical formula for combinations) that there are also only two possible ways of getting a $U$ of 1, four ways of getting a $U$ of 2, six ways of getting a $U$ of 3, 10 ways for a $U$ of 4, and 14 ways for a $U$ of 5. In sum, there are 38 possible ways of producing a $U$ value of 5 or less by chance alone. Dividing 38 by 462 gives a probability of .082 of this result under the null hypothesis. Thus, for a two-tailed prediction, with sample sizes of 5 and 6, a $U$ value of 5 is not significant at the $p = .05$ level of significance.

It is interesting to note that if we had made a one-tailed prediction, that is, specifically predicted in advance that Alba's students would improve less than Bolt's, then there would be only 19 ways of getting 5 or less, and the probability by chance would be .041. With this one-tailed prediction, a $U$ value of 5 would now produce a significant difference at the significance level of $p = .05$.

However, one does not normally need to work out the probability for the calculated $U$ values. It can be looked up in statistical tables. Alternatively, a software package will give the probability of a $U$ value, which can be compared to the chosen significance level.

The Mann–Whitney $U$ test is a useful test of small samples (Mann and Whitney, 1947, gave tables of the probability of $U$ for samples of $n_1$ and $n_2$ up to 8), but with large sample sizes ($n_1$ and $n_2$ both greater than 20), then the $U$ distribution tends to a normal distribution with a mean of

$$\frac{n_1 n_2}{2}$$

and standard deviation of

$$\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}.$$

With large samples, rather than the $U$ test, a $z$ score can be calculated and the probability looked up in the standard normal distribution tables.

The accuracy of the Mann–Whitney $U$ test is reduced the more tied ranks there are, and so where there are a large number of tied ranks, a correction to the test is necessary. Indeed, in this case, it is often worth considering whether a more precise measure of the dependent variable is called for.

*Perry R. Hinton*

***See also*** Dependent Variable; Independent Variable; Nonparametric Statistics; Null Hypothesis; One-Tailed Test; Sample; Significance Level, Concept of; Significance Level, Interpretation and Construction; Two-Tailed Test; Wilcoxon Rank Sum Test; $z$ Score

### Further Readings

Conover, W. J. (1999). *Practical nonparametric statistics* (3rd ed.). New York: Wiley.

Daniel, W. W. (1990). *Applied nonparametric statistics* (2nd ed.). Boston: PWS-KENT.

Gibbons, J. D., & Chakraborti, S. (2003). *Nonparametric statistical inference* (4th ed.). New York: Marcel Dekker.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics, 18,* 50–60.

Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioural sciences.* New York: McGraw-Hill.

## Margin of Error

In the popular media, the margin of error is the most frequently quoted measure of statistical accuracy for a sample estimate of a population parameter. Based on the conventional definition of the measure, the difference between the estimate and the targeted parameter should be bounded by the margin of error 95% of the time. Thus, only 1 in 20 surveys or studies should lead to a result in which the actual estimation error exceeds the margin of error.

Technically, the margin of error is defined as the radius or the half-width of a symmetric confidence interval. To formalize this definition, suppose that the targeted population parameter is denoted by $\theta$. Let $\hat{\theta}$ represent an estimator of $\theta$ based on the sample data. Let $SD(\hat{\theta})$ denote the standard deviation of $\hat{\theta}$ (if known) or an estimator of the standard deviation (if unknown). $SD(\hat{\theta})$ is often referred to as the *standard error.*

Suppose that the sampling distribution of the standardized statistic

$$\frac{(\hat{\theta} - \theta)}{SD(\hat{\theta})}$$

is symmetric about zero. Let $Q(0.975)$ denote the 97.5th percentile of this distribution. (Note that the 2.5th percentile of the distribution would then be given by $-Q(0.975)$.) A symmetric 95% confidence interval for $\theta$ is defined as $\hat{\theta} \pm Q(0.975)\, SD(\hat{\theta})$. The half-width or radius of such an interval, $ME = Q(0.975)SD(\hat{\theta})$, defines the conventional margin of error, which is implicitly based on a 95% confidence level.

A more general definition of the margin of error is based on an arbitrary confidence level. Suppose $100(1-\alpha)\%$ represents the confidence level corresponding to a selected value of $\alpha$ between 0 and 1. Let $Q(1-\frac{\alpha}{2})$ denote the $100(1-\frac{\alpha}{2})$th percentile of the sampling distribution of $\frac{(\hat{\theta}-\theta)}{SD(\hat{\theta})}$. A symmetric $100(1-\alpha)\%$ confidence interval is given by

$$\hat{\theta} \pm Q\left(1-\frac{\alpha}{2}\right) SD(\hat{\theta}),$$

leading to a margin of error of

$$ME(\alpha) = Q\left(1-\frac{\alpha}{2}\right) SD(\hat{\theta}).$$

The size of the margin of error is based on three factors: (1) the size of the sample, (2) the variability of the data being sampled from the population, and (3) the confidence level (assuming that the conventional 95% level is not employed). The sample size and the population variability are both reflected in the standard error of the estimator, $SD(\hat{\theta})$, which decreases as the sample size increases and grows in accordance with the dispersion of the population data. The confidence level is represented by the percentile of the sampling distribution, $Q(1-\frac{\alpha}{2})$. This percentile becomes larger as $\alpha$ is decreased and the corresponding confidence level $100(1-\alpha)\%$ is increased.

A common problem in research design is sample size determination. In estimating a parameter $\theta$, an investigator often wishes to determine the sample size $n$ required to ensure that the margin of error does not exceed some predetermined bound $B$; that is, to find the $n$ that will ensure $ME(\alpha) \leq B$. Solving this problem requires specifying the confidence level as well as quantifying the population variability. The latter is often accomplished by relying on data from pilot or preliminary studies, or from prior studies that investigate similar phenomena. In some instances (such as when the parameter of interest is a proportion), an upper bound can be placed on the population variability. The use of such a bound results in a conservative sample size determination; that is, the resulting $n$ is at least as large as (and possibly larger than) the sample size actually required to achieve the desired objective.

Two of the most basic problems in statistical inference consist of estimating a population mean and estimating a population proportion under

random sampling with replacement. The margins of error for these problems are presented in the next sections.

### Margin of Error for Means

Assume that a random sample of size $n$ is drawn from a quantitative population with mean $\mu$ and standard deviation $\sigma$. Let $\overline{x}$ denote the sample mean. The standard error of $\overline{x}$ is then given by

$$\frac{\sigma}{\sqrt{n}}.$$

Assume that either (a) the population data may be viewed as normally distributed or (b) the sample size is "large" (typically, 30 or greater). The sampling distribution of the standardized statistic

$$\frac{(\overline{x}-\mu)}{\left(\frac{\sigma}{\sqrt{n}}\right)}$$

then corresponds to the standard normal distribution, either exactly (under normality of the population) or approximately (in a large sample setting, by virtue of the central limit theorem). Let $Z(1-\frac{\alpha}{2})$ denote the $100(1-\frac{\alpha}{2})$th percentile of this distribution. The margin of error for $\overline{x}$ is then given by

$$ME(\alpha) \;=\; Z\left(1-\frac{\alpha}{2}\right)\left(\frac{\sigma}{\sqrt{n}}\right).$$

With the conventional 95% confidence level, $\alpha = 0.05$ and $Z(0.975) = 1.96 \approx 2$, leading to $ME = 2\,(\frac{\sigma}{\sqrt{n}})$.

In research design, the sample size needed to ensure that the margin of error does not exceed some bound $B$ is determined by finding the smallest $n$ that will ensure

$$n \geq \left(Z\left(1-\frac{\alpha}{2}\right)\frac{\sigma}{B}\right)^2.$$

In instances in which the population standard deviation $\sigma$ cannot be estimated based on data collected from earlier studies, a conservative approximation of $\sigma$ can be made by taking one fourth the plausible range of the variable of interest (i.e., $(\frac{1}{4})$ [maximum–minimum]).

The preceding definition for the margin of error assumes that the standard deviation $\sigma$ is known, an assumption that is unrealistic in practice. If $\sigma$ is unknown, it must be estimated by the sample standard deviation $s$. In this case, the margin of error is based on the sampling distribution of the statistic

$$\frac{(\overline{x}-\mu)}{\left(\frac{s}{\sqrt{n}}\right)}.$$

This sampling distribution corresponds to the Student's $t$ distribution, either exactly (under normality of the population) or approximately (in a large sample setting). If $T_{df}(1-\frac{\alpha}{2})$ denotes the $100(1-\frac{\alpha}{2})$th percentile of the $t$ distribution with $df = n-1$ degrees of freedom, then the margin of error for $\overline{x}$ is given by

$$ME(\alpha) \;=\; T_{df}\left(1-\frac{\alpha}{2}\right)\left(\frac{s}{\sqrt{n}}\right).$$

### Margin of Error for Proportions

Assume that a random sample of size $n$ is drawn from a qualitative population where $\pi$ denotes a proportion based on a characteristic of interest. Let $p$ denote the sample proportion. The standard error of $p$ is then given by

$$\sqrt{\frac{\pi(1-\pi)}{n}}.$$

Here, $\sqrt{\pi(1-\pi)}$ represents the standard deviation of the binary (0/1) population data, in which each object is dichotomized according to whether it exhibits the characteristic in question. The sample version of the standard error is

$$\sqrt{\frac{p(1-p)}{n}}.$$

Assume that the sample size is "large" (typically, such that $n\pi$ and $n(1-\pi)$ are both at least 10). The sampling distribution of the standardized statistic

$$\frac{(p-\pi)}{\left(\sqrt{\frac{\pi(1-\pi)}{n}}\right)}$$

can then be approximated by the standard normal distribution. The margin of error for $\overline{x}$ is then given by

$$ME(\alpha) = Z\left(1-\frac{\alpha}{2}\right)\left(\sqrt{\frac{p(1-p)}{n}}\right).$$

With the conventional 95% confidence level, $ME = 2\left(\sqrt{\frac{p(1-p)}{n}}\right)$.

In research design for sample size determination, for applications in which no data exist from previous studies for estimating the proportion of interest, the computation is often based on bounding the population standard deviation $\sqrt{\pi(1-\pi)}$. Using calculus, it can be shown that this quantity achieves a maximum value of 1/2 when the proportion $\pi$ is 1/2. The maximum margin of error is then defined as

$$ME_{MAX}(\alpha) = Z\left(1-\frac{\alpha}{2}\right)\left(\frac{1}{2\sqrt{n}}\right).$$

The sample size needed to ensure that this margin of error does not exceed some bound $B$ is determined by finding the smallest $n$ that will ensure $n \geq \left(\frac{Z(1-\frac{\alpha}{2})}{2B}\right)^2$.

Often in public opinion polls and other surveys, a number of population proportions are estimated, and a single margin of error is quoted for the entire set of estimates. This is generally accomplished with the preceding maximum margin of error $ME_{\mathrm{MAX}}(\alpha)$, which is guaranteed to be at least as large as the margin of error for any of the individual estimates in the set. When the conventional 95% confidence level is employed, the maximum margin of error has a particularly simple form:

$$ME_{\mathrm{MAX}} = \frac{1}{\sqrt{n}}.$$

National opinion polls are often based on a sample size of roughly 1,100 participants, leading to the margin of error $ME_{\mathrm{MAX}} \approx 0.03$, or 3 percentage points.

## Finite Population Correction

The preceding formulas for margins of error are based on the assumption that the sample is drawn from the population at random with replacement. If the sample is drawn at random without replacement, and the sampling fraction is relatively high (e.g., 5% or more), the formulas should be adjusted by a *finite population correction* (fpc). If $N$ denotes the size of the population, this correction is given as

$$\mathrm{fpc} = \sqrt{\frac{(N-n)}{(N-1)}}.$$

Employing this correction has the effect of reducing the margin of error. As $n$ approaches $N$, the fpc becomes smaller. When $N = n$, the entire population is sampled. In this case, the fpc and the margin of error are zero because the sample estimate is equal to the population parameter and there is no estimation error. In practice, the fpc is generally ignored because the size of the sample is usually small relative to the size of the population.

## Generalization to Two Parameters

The preceding definitions for the margin of error are easily generalized to settings in which two targeted population parameters are of interest, say $\theta_1$ and $\theta_2$. Two parameters are often compared by estimating their difference, say $\theta_1 - \theta_2$. Let $\hat{\theta}_1$ and $\hat{\theta}_2$ represent estimators of $\theta_1$ and $\theta_2$ based on the sample data. Let $SD(\hat{\theta}_1 - \hat{\theta}_2)$ denote the standard error of the difference in the estimators $\hat{\theta}_1 - \hat{\theta}_2$.

Confidence intervals for $\theta_1 - \theta_2$ may be based on the standardized statistic

$$\frac{(\hat{\theta}_1 - \hat{\theta}_2) - (\theta_1 - \theta_2)}{SD(\hat{\theta}_1 - \hat{\theta}_2)}.$$

As before, suppose that the sampling distribution of this statistic is symmetric about zero, and let $Q(0.975)$ denote the 97.5th percentile of this distribution. A symmetric 95% *confidence interval* for $\theta$ is defined as

$$(\hat{\theta}_1 - \hat{\theta}_2) \pm Q(0.975)SD(\hat{\theta}_1 - \hat{\theta}_2).$$

The half-width or radius of such an interval, $ME = Q(0.975)SD(\hat{\theta}_1 - \hat{\theta}_2)$, defines the conventional margin of error for the difference $\hat{\theta}_1 - \hat{\theta}_2$, which is based on a 95% confidence level. The more general definition based on an arbitrary

confidence level $100(1-\alpha)\%$ may be obtained by replacing the 97.5th percentile $Q(0.975)$ with the $100(1-\frac{\alpha}{2})$th percentile $Q(1-\frac{\alpha}{2})$, leading to

$$ME(\alpha) = Q\left(1-\frac{\alpha}{2}\right)SD(\hat{\theta}_1-\hat{\theta}_2).$$

*Joseph E. Cavanaugh and Eric D. Foster*

***See also*** Confidence Intervals; Estimation; Sample Size Planning; Standard Deviation; Standard Error of Estimate; Standard Error of the Mean; Student's *t* Test; *z* Distribution

**Further Readings**

Lohr, S. L. (1999). *Sampling: Design and analysis*. Pacific Grove, CA: Duxbury Press.
Utts, J. M. (2005). *Seeing through statistics* (3rd ed.). Belmont, CA: Thomson Brooks/Cole.
Utts, J. M., & Heckard, R. F. (2007). *Mind on statistics* (3rd ed.). Belmont, CA: Thomson Brooks/Cole.

# Markov Chains

The topic of Markov chains is a well-developed topic in probability. There are many fine expositions of Markov chains (e.g., Bremaud, 2008; Feller, 1968; Hoel, Port, & Stone, 1972; Kemeny & Snell, 1960). Those expositions and others have informed this concise entry on Markov chains, which is not intended to exhaust the topic of Markov chains. The topic is just too capacious for that. This entry provides an exposition of a judicious sampling of the major ideas, concepts, and methods regarding the topic.

## Andrei Andreevich Markov

Andrei Andreevich Markov (1856–1922) formulated the seminal concept in the field of probability later known as the Markov chain. Markov was an eminent Russian mathematician who served as a professor in the Academy of Sciences at the University of St. Petersburg. One of Markov's teachers was Pafnuty Chebyshev, a noted mathematician who formulated the famous inequality termed the *Chebyschev inequality,* which is extensively used in probability and statistics. Markov was the first person to provide a clear proof of the central limit theorem, a pivotal theorem in probability and statistics that indicates that the sum of a large number of independent random variables is asymptotically distributed as a normal distribution.

Markov was a well-trained mathematician who after 1900 emphasized inquiry in probability. After studying sequences of independent chance events, he became interested in sequences of mutually dependent events. This inquiry led to the creation of Markov chains.

## Sequences of Chance Events

Markov chains are sequences of chance events. A series of flips of a fair coin is a typical sequence of chance events. Each coin flip has two possible outcomes: Either a head ($H$) appears or a tail ($T$) appears. With a fair coin, a head will appear with a probability ($p$) of 1/2 and a tail will appear with a probability of 1/2.

Successive coin flips are independent of each other in the sense that the probability of a head or a tail on the first flip does not affect the probability of a head or a tail on the second flip. In the case of the sequence $HT$, the $p(HT) = p(H) \times p(T) = 1/2 \times 1/2 = 1/4$. Many sequences of chance events are composed of independent chance events such as coin flips or dice throws. However, some sequences of chance events are not composed of independent chance events. Some sequences of chance events are composed of events whose occurrences are influenced by prior chance events. Markov chains are such sequences of chance events.

As an example of a sequence of chance events that involves interdependence of events, let us consider a sequence of events $E_1, E_2, E_3, E_4$ such that the probability of any of the events after $E_1$ is a function of the prior event. Instead of interpreting the $p(E_1, E_2, E_3, E_4)$ as the product of probabilities of independent events, $p(E_1, E_2, E_3, E_4)$ is the product of an initial event probability and the conditional probabilities of successive events. From this perspective,

$$p(E_1, E_2, E_3, E_4) = p(E_1) \times p(E_2|E_1) \times p(E_3|E_2) \times p(E_4|E_3).$$

Such a sequence of events is a Markov chain.

Let us consider a sequence of chance events $E_1$, $E_2$, $E_3, \ldots, E_j, \ldots, E_n$. If $p(E_j|E_{j-1}) = p(E_j|E_{j-1},$

$E_{j-2}, \ldots, E_1$), then the sequence of chance events is a Markov chain. For a more formal definition, if $X_1, X_2, \ldots, X_n$ are random variables and if $p(X_n = k_n | X_{n-1} = k_{n-1}) = p(X_n = k_n | X_{n-1} = k_{n-1}, \ldots, X_2 = k_2, X_1 = k_2)$, then $X_1, X_2, \ldots, X_n$ form a Markov chain.

Conditional probabilities interrelating events are important in defining a Markov chain. Common in expositions of Markov chains, conditional probabilities interrelating events are termed *transition probabilities interrelating states,* events are termed *states,* and a set of states is often termed a *system* or a *state space*. The states in a Markov chain are either finite or countably infinite; this entry will feature systems or state spaces whose states are finite in number.

A matrix of *transition* probabilities is used to represent the interstate transitions possible for a Markov chain. As an example, consider the following system of states, $S_1$, $S_2$, and $S_3$, with the following matrix of transition probabilities, $\mathbf{P}_1$.

$$\mathbf{P}_1 = \begin{array}{c} \\ S_1 \\ S_2 \\ S_3 \end{array} \begin{array}{c} \begin{array}{ccc} S_1 & S_2 & S_3 \end{array} \\ \begin{bmatrix} .1 & .8 & .1 \\ .5 & .3 & .2 \\ .1 & .7 & .2 \end{bmatrix} \end{array}$$

Using the matrix $\mathbf{P}_1$, one can see that the transition probability of entering $S_1$ given that one is in $S_2$ is .8. That same probability is represented as $p_{12}$.

## Features of Markov Chains and Their States

There are various types of states in a Markov chain. Some types of states in a Markov chain relate to the degree to which states recur over time. A *recurrent state* is one that will return to itself before an infinite number of steps with probability 1. An *absorbing state i* is a recurrent state for which $p_{ii} = 1$ and $p_{ij} = 0$ for $i \neq j$. In other words, if it is not possible to leave a given state, then that state is an absorbing state. Second, a state is *transient* if it is not recurrent. In other words, if the probability that a state will occur again before an infinite number of steps is less than 1, then that state is transient.

To illustrate the attributes of absorbing and transient states, let us consider the following example. A type of job certification involves a test with three resulting states. With state $S_1$, one fails the test with a failing score and then maintains that failure status. With state $S_2$, one passes the test with a low pass score that is inadequate for certification. One then takes the test again. Either one attains state $S_2$ again with a probability of .5 or one passes the test with a high pass score and reaches state $S_3$ with a probability of .5. With state $S_3$, one passes the test with a high pass score that warrants job certification.

$$\mathbf{P}_2 = \begin{array}{c} \\ S_1 \\ S_2 \\ S_3 \end{array} \begin{array}{c} \begin{array}{ccc} S_1 & S_2 & S_3 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 \\ 0 & .5 & .5 \\ 0 & 0 & 1 \end{bmatrix} \end{array}$$

$\mathbf{P}_2$ indicates the transition probabilities among the three states of the job certification process. From an examination of $\mathbf{P}_2$, state $S_1$ is an absorbing state because $p_{11} = 1$ and $p_{1j} = 0$ for $1 \neq j$, and state $S_3$ is an absorbing state because $p_{33} = 1$ and $p_{3j} = 0$ for $3 \neq j$. However, state $S_2$ is a transient state because there is a nonzero probability that the state will never be reached again.

To illustrate that state $S_2$ will never be reached again, let us examine what happens as the successive steps in the Markov chain occur. $\mathbf{P}_2$ indicates the transition probabilities among the three states of the job certification process. $\mathbf{P}_2^2$ indicates the transition probabilities after two steps in the process.

$$\mathbf{P}_2^2 = \begin{array}{c} \\ S_1 \\ S_2 \\ S_3 \end{array} \begin{array}{c} \begin{array}{ccc} S_1 & S_2 & S_3 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 \\ 0 & .25 & .75 \\ 0 & 0 & 1 \end{bmatrix} \end{array}$$

After two steps, $p_{22}$ has decreased to .25 and $p_{23}$ has increased to .75. $\mathbf{P}_2^3$ indicates the transition probabilities after two steps in the process.

$$\mathbf{P}_2^3 = \begin{array}{c} \\ S_1 \\ S_2 \\ S_3 \end{array} \begin{array}{c} \begin{array}{ccc} S_1 & S_2 & S_3 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 \\ 0 & .06 & .94 \\ 0 & 0 & 1 \end{bmatrix} \end{array}$$

After three steps, $p_{22}$ has further decreased to .06 and $p_{23}$ has further increased to .94. One could surmise that $p_{22}$ will further decrease and $p_{23}$ will further decrease as the number of steps increases. $\mathbf{P}_2^2$ and $\mathbf{P}_2^3$ suggest that those receiving an intermediate score will pass the test with successive attempts.

Generalizing to $n$ steps, $\mathbf{P}_2^n$ indicates the transition probabilities after $n$ steps in the process.

$$\mathbf{P}_2^n = \begin{array}{c} \\ S_1 \\ S_2 \\ S_3 \end{array} \begin{array}{c} \begin{array}{ccc} S_1 & S_2 & S_3 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 \\ 0 & (.5)^n & 1-(.5)^n \\ 0 & 0 & 1 \end{bmatrix} \end{array}$$

As $n$ increases, $p_{22}$ will approach 0 and $p_{23}$ will approach 1. From an examination of $\mathbf{P}_2^n$, it is clear that those receiving an intermediate score will inexorably pass the test with successive attempts at taking the test. As the number of steps in the process increases, individuals will either have failed the job certification procedure or have passed the job certification procedure.

Additional types of states in a Markov chain relate to the degree to which states can be reached from other states in the chain. A state $i$ is *accessible* from a different state $j$ if there is a nonzero probability that state $j$ can be reached from state $i$ some time in the future. Second, a state $i$ *communicates* with state $j$ if state $i$ is accessible from state $j$ and state $j$ is accessible from state $i$. A set of states in a Markov chain is a *communicating* class if every pair of states in the set communicates with each other. A communicating class is *closed* if it is not possible to reach a state outside the class from a state within the class. A Markov chain is *irreducible* if any state within the chain is accessible from any other state in the chain. An irreducible Markov chain is also termed an *ergodic* Markov chain.

To illustrate these attributes, let us consider the following Markov chain and its transition probability matrix. This hypothetical chain describes the transition from the IQ of the parent to the IQ of a child of the parent. Let the states be three levels of IQ: high, intermediate, and low. $\mathbf{P}_3$ is the transition matrix for this Markov chain.

$$\mathbf{P}_3 = \begin{array}{c} \\ \textit{Parental IQ} \\ S_1 \ \text{(low)} \\ S_2 \ \text{(intermediate)} \\ S_3 \ \text{(high)} \end{array} \begin{array}{c} \textit{Child IQ} \\ \begin{array}{ccc} S_1 & S_2 & S_3 \\ \text{(low)} & \text{(intermediate)} & \text{(high)} \end{array} \\ \begin{bmatrix} .6 & .3 & .1 \\ .2 & .6 & .2 \\ .1 & .3 & .6 \end{bmatrix} \end{array}$$

The Markov chain indicated by $\mathbf{P}_3$ reflects the view that the IQ of the parent tends to determine the IQ of the child.

All the states in $\mathbf{P}_3$ are accessible from each other, and they all communicate with each other. Thus the states in $\mathbf{P}_3$ form a communicating class. In addition, this Markov chain is irreducible and ergodic.

But what would be the probabilities of the three levels of parental IQ if this Markov chain would proceed for many steps? To determine these probabilities, also called *stationary* probabilities, one needs to use an important finding in the study of Markov chains.

Let us assume that we have a transition probability matrix, $\mathbf{P}$, which, for some $n$, $\mathbf{P}^n$ has only nonzero entries. Then $\mathbf{P}$ is termed a *regular* matrix. Then two prominent theorems follow.

The first theorem addresses the increasing similarity of rows in successive powers of $\mathbf{P}$. If $\mathbf{P}$ is a regular transition probability matrix, then $\mathbf{P}^n$ becomes increasing more similar to a probability matrix $\prod$ with each row of $\prod$ being the same probability vector $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_k)$ and with the components of $\boldsymbol{\pi}$ being positive.

The second theorem indicates the equation that permits the computation of stationary probabilities. If $\mathbf{P}$ is a regular transition probability matrix, then the row vector $\boldsymbol{\pi}$ is the unique vector of probabilities that satisfies the equation $\boldsymbol{\pi} = \boldsymbol{\pi} \cdot \mathbf{P}$. This theorem is one of the more important theorems in the study of Markov chains.

With the equation $\boldsymbol{\pi} = \boldsymbol{\pi} \cdot \mathbf{P}$, we can determine the stationary probabilities for $\mathbf{P}_3$. Let $\boldsymbol{\pi}_3 = (\pi_{31}, \pi_{32}, \pi_{33})$ be the vector of stationary probabilities for this three-state Markov chain associated with $\mathbf{P}_3$. Then $\boldsymbol{\pi}_3 = \boldsymbol{\pi}_3 \cdot \mathbf{P}_3$.

$$(\pi_{31}, \pi_{32}, \pi_{33}) = (\pi_{31}, \pi_{32}, \pi_{33}) \cdot \begin{bmatrix} .6 & .3 & .1 \\ .2 & .6 & .2 \\ .1 & .3 & .6 \end{bmatrix}$$

$$= (.6\pi_{31} + .2\pi_{32} + .1\pi_{33}, .3\pi_{31} + .6\pi_{32} + .3\pi_{33}, .1\pi_{31} + .2\pi_{32} + .6\pi_{33})$$

This results in three equations:

(1) $\pi_{31} = .6\pi_{31} + .2\pi_{32} + .1\pi_{33}$

(2) $\pi_{32} = .3\pi_{31} + .6\pi_{32} + .3\pi_{333}$

(3) $\pi_{33} = .1\pi_{31} + .2\pi_{32} + .6\pi_{33}$

Along with those three equations is the additional equation:

(4) $\pi_{31} + \pi_{32} + \pi_{33} = 1.$

An arithmetic manipulation of these four equations results in numerical solutions for the three unknowns: $\pi_{31} = 2/7 = .286$, $\pi_{32} = 3/7 = .428$, and $\pi_{33} = 2/52 = 2/7 = .286$. These three stationary probabilities indicate that there would be a modest trend toward intermediate IQs over many generations.

Let us consider another Markov chain and its transition probability matrix. This hypothetical chain also describes the transition from the IQ of the parent to the IQ of a child of the parent. Let the states again be three levels of IQ: high, intermediate, and low. $\mathbf{P}_4$ is the transition matrix for this Markov chain.

| | | *Child IQ* | | |
|---|---|---|---|---|
| | | $S_1$ | $S_2$ | $S_3$ |
| *Parental IQ* | | (low) | (intermediate) | (high) |
| $S_1$ | (low) | .4 | .4 | .2 |
| $P_4 = S_2$ | (intermediate) | .3 | .4 | .3 |
| $S_3$ | (high) | .2 | .4 | .4 |

The Markov chain indicated by $\mathbf{P}_4$ reflects the view that the IQ of the parent weakly influences the IQ of the child.

What would be the probabilities of the three levels of parental IQ if this Markov chain would proceed for many steps? To determine these probabilities, also called stationary probabilities, one needs to solve the equation $\boldsymbol{\pi}_4 = \boldsymbol{\pi}_4 \cdot \mathbf{P}_4$ with $\boldsymbol{\pi}_4 = (\pi_{41}, \pi_{42}, \pi_{43})$, the vector of stationary probabilities for this three-state Markov chain.

$$\begin{bmatrix} .4 & .4 & .2 \\ .3 & .4 & .3 \\ .2 & .4 & .4 \end{bmatrix}$$

$$= (.4\pi_{41} + .3\pi_{42} + .2\pi_{43}, .4\pi_{41} + .4\pi_{42} + .4\pi_{43}, .2\pi_{41} + .3\pi_{42} + .4\pi_{43})$$

This results in three equations:

(1) $\pi_{41} = .4\pi_{41} + .3\pi_{42} + .2\pi_{43}$

(2) $\pi_{42} = .4\pi_{41} + .4\pi_{42} + .4\pi_{43}$

(3) $\pi_{43} = .2\pi_{41} + .3\pi_{42} + .4\pi_{43}$

Along with those three equations is the additional equation:

(4) $\pi_{41} + \pi_{42} + \pi_{43} = 1.$

An arithmetic manipulation of these four equations results in numerical solutions for the three unknowns: $\pi_{41} = 3/10 = .3$, $\pi_{42} = 2/5 = .4$, and $\pi_{43} = 3/10 = .3$. These three stationary probabilities indicate that there would be a weak trend toward intermediate IQs over many generations with this Markov chain but not as strong a trend as with the prior Markov chain depicted with $\mathbf{P}_3$.

Not all Markov chains have unique stationary probabilities. Let us consider the matrix of transition probabilities, $\mathbf{P}_2$, with two absorbing states.

$$\begin{array}{r} \\ S_1 \\ P_2 = S_2 \\ S_3 \end{array} \begin{array}{c} \begin{array}{ccc} S_1 & S_2 & S_3 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 \\ 0 & .5 & .5 \\ 0 & 0 & 1 \end{bmatrix} \end{array}$$

To determine the stationary probabilities, one needs to solve the equation: $\boldsymbol{\pi}_2 = \boldsymbol{\pi}_2 \cdot \mathbf{P}_2$ with $\boldsymbol{\pi}_2 = (\pi_{21}, \pi_{22}, \pi_{23})$, the vector of stationary probabilities for this three-state Markov chain.

$$(\pi_{21}, \pi_{22}, \pi_{23}) = (\pi_{21}, \pi_{22}, \pi_{23}) \cdot \begin{bmatrix} 1 & 0 & 0 \\ 0 & .5 & .5 \\ 0 & 0 & 1 \end{bmatrix}$$

$$= (\pi_{21}, .5\pi_{22} + .5\pi_{23}, .5\pi_{22} + .5\pi_{23})$$

This results in three equations:

1. $\pi_{21} = \pi_{21}$
2. $\pi_{22} = \ + .5\pi_{22}$
3. $\pi_{23} = \ + .5\pi_{22} + \pi_{23}$

Along with those three equations is the additional equation:

4. $\pi_{21} + \pi_{22} + \pi_{23} = 1.$

An arithmetic manipulation of these 4 equations results in numerical solutions for the three unknowns: $\pi_{21} = 1 - \pi_{23}$, $\pi_{22} = 0$, and $\pi_{23} = 1 - \pi_{21}$. These three stationary probabilities indicate that there could be an infinite number of values for $\pi_{21}$ and $\pi_{23}$.

Additional types of states in a Markov chain relate to the rates at which states in the Markov chain return to themselves over time. If there is a return to state *i* starting from state *i* after every three steps, then state *i* has a *period* of 3. In general, if there is a return to a given state in a Markov chain after every *t* steps if one starts with that state, then that state is periodic with period *t*. A state is recurrent or persistent if there is a probability of 1 that the state will be reached again.

To illustrate these attributes, let us consider the following example. Let $\mathbf{P}_5$ be a matrix of transition probabilities for a Markov chain with three absorbing states.

$$\mathbf{P}_5 = \begin{array}{c} \\ S_1 \\ S_2 \\ S_3 \end{array} \begin{array}{c} \begin{array}{ccc} S_1 & S_2 & S_3 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \end{array}$$

After one step in the Markov chain, $\mathbf{P}_5^2$ indicates the resulting transition probabilities.

$$\mathbf{P}_5^2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$\mathbf{P}_5^3 =$ is the identity matrix. After two steps in the Markov chain, $\mathbf{P}_5^3$ indicates the resulting transition probabilities.

$$\mathbf{P}_5^3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

$\mathbf{P}_5^3 = \mathbf{P}_5$. One could then generalize to the following equality: $\mathbf{P}_5^{2n+1} = \mathbf{P}_5$ with $n =$ any natural number. Regarding the attributes of periodicity or recurrence, state $S_{11}$ is recurrent at each step and states $S_{13}$ and $S_{32}$ are periodic with period 2.

One type of Markov chain that is commonplace among expositions of Markov chains is that of the random walk. Imagine a person named Harry, a frequent visitor to a bar, planning to leave the bar, represented as state $S_2$. From the bar, Harry could start walking to his home and return to the bar, $S_1$, or walk to a park, state $S_2$, that is halfway between the bar and his home. If Harry is at the park, state $S_2$, then he could start walking to his home and return to the park or walk to his home, state $S_3$. If Harry is at his home, then Harry could remain at his home or walk to the park. $\mathbf{P}_6$ is the matrix of transition probabilities that relates to this random walk.

$$\mathbf{P}_6 = \begin{array}{c} \\ S_1 \text{ (bar)} \\ S_2 \text{ (park)} \\ S_3 \text{ (home)} \end{array} \begin{array}{c} \begin{array}{ccc} S_1 & S_2 & S_3 \end{array} \\ \begin{bmatrix} .5 & .5 & 0 \\ .5 & 0 & .5 \\ 0 & .5 & .5 \end{bmatrix} \end{array}$$

To determine the stationary probabilities, one needs to solve the equation: $\boldsymbol{\pi}_6 = \boldsymbol{\pi}_6 \cdot \mathbf{P}_6$ with $\boldsymbol{\pi}_6 = (\pi_{61}, \pi_{62}, \pi_{63})$, the vector of stationary probabilities for this three-state Markov chain.

$$(\pi_{61}, \pi_{62}, \pi_{63}) = (\pi_{61}, \pi_{62}, \pi_{63}) \cdot \begin{bmatrix} .5 & .5 & 0 \\ .5 & 0 & .5 \\ 0 & .5 & .5 \end{bmatrix}$$

$$= (.5\pi_{61} + .5\pi_{62},\ .5\pi_{61} + .5\pi_{63},\ .5\pi_{62} + .5\pi_{63})$$

This results in three equations:

1. $\pi_{61} = .5\pi_{61} + .5\pi_{62}$
2. $\pi_{62} = .5\pi_{61} + .5\pi_{63}$
3. $\pi_{63} = .5\pi_{62} + .5\pi_{63}$

Along with those three equations is the additional equation

4. $\pi_{61} + \pi_{62} + \pi_{63} = 1$.

An arithmetic manipulation of these four equations results in numerical solutions for the three unknowns: $\pi_{61} = \pi_{62} = \pi_{63} = 1/3 = .33$. These stationary probabilities indicate that Harry would be at any of the three locations with equal likelihood after many steps in the random walk.2

### Conclusion

If there is a sequence of random events such that a future event is dependent only on the present event and not on past events, then the sequence is likely a Markov chain, and the work of Markov and others may be used to extract useful information from an analysis of the sequence. The topic of the Markov chain has become one of the most captivating, generative, and useful topics in probability and statistics.

*William M. Bart and Thomas Bart*

***See also*** Matrix Algebra; Probability, Laws of

#### Further Readings

Bremaud, P. (2008). *Markov chains*. New York: Springer.
Feller, W. (1968). *An introduction to probability theory and its applications: Volume 1* (3rd ed.). New York: Wiley.
Hoel, P., Port, S., & Stone, C. (1972). *Introduction to stochastic processes*. Boston: Houghton Mifflin.
Kemeny, J., & Snell, J. (1960). *Finite Markov chains*. Princeton, NJ: D. Van Nostrand.

## Matching

The term *matching* refers to the procedure of finding for a sample unit other units in the sample that are closest in terms of observable characteristics. The units selected are usually referred to as *matches,* and after repeating this procedure for all units (or a subgroup of them), the resulting subsample of units is called the *matched sample*. This idea is typically implemented across subgroups of a given sample, that is, for each unit in one subgroup, matches are found among units of another subgroup. A matching procedure requires defining a notion of distance, selecting the number of matches to be found, and deciding whether units will be used multiple times as a potential match. In applications, matching is commonly used as a preliminary step in the construction of a matched sample, that is, a sample of observations that are similar in terms of observed characteristics, and then some statistical procedure is computed with this subsample. Typically, the term *matching estimator* refers to the case when the statistical procedure of interest is a point estimator, such as the sample mean. The idea of matching is usually employed in the context of observational studies, in which it is assumed that selection into treatment, if present, is based on observable characteristics. More generally, under appropriate assumptions, matching may be used as a way of reducing variability in estimation, combining databases from different sources, dealing with missing data, and designing sampling strategies, among other possibilities. Finally, in the econometrics literature, the term *matching* is sometimes used more broadly to refer to a class of estimators that exploit the idea of selection on observables in the context of program evaluation. This entry focuses on the implementation of and statistical inference procedures for matching.

### Description and Implementation

A natural way of describing matching formally is in the context of the classical potential outcomes model. To describe this model, suppose that a random sample of size $n$ is available from a large population, which is represented by the collection of random variables $(Y_i, T_i, X_i)$, $i = 1, 2, \ldots, n$, where $T_i \in \{0, 1\}$,

$$Y_i = \begin{cases} Y_{0i} & \text{if} \quad T_i = 0 \\ Y_{1i} & \text{if} \quad T_i = 1 \end{cases}$$

and $X_i$ represents a (possibly high-dimensional) vector of observed characteristics. This model aims to capture the idea that while the set of characteristics $X_i$ is observed for all units, only one of the two random variables $(Y_{0i}, Y_{1i})$ is observed for each unit, depending on the value of $T_i$. The underlying random variables $Y_{0i}$ and $Y_{1i}$ are usually referred to as potential outcomes because they represent the two potential states for each unit.

For example, this model is routinely used in the program evaluation literature, where $T_i$ represents treatment status and $Y_{0i}$ and $Y_{1i}$ represent outcomes without and with treatment, respectively. In most applications the goal is to establish statistical inference for some characteristic of the distribution of the potential outcomes such as the mean or quantiles. However, using the available sample directly to establish inference may lead to important biases in the estimation whenever units have selected into one of the two possible groups ($T_i = 0$ or $T_i = 1$). As a consequence, researchers often assume that the selection process, if present, is based on observable characteristics. This idea is formalized by the so-called *conditional independence assumption:* conditionally on $X_i$, the random variables $(Y_{0i}, Y_{1i})$ are independent of $T_i$. In other words, under this assumption, units having the same observable characteristics $X_i$ are assigned to each of the two groups ($T_i = 0$ or $T_i = 1$) independently of their potential gains, captured by $(Y_{0i}, Y_{1i})$. Thus, this assumption imposes random treatment assignment conditional on $X_i$. This model also assumes some form of *overlap* or *common support:* For some $c > 0$, $c \leq \mathbb{P}(T_i = 1|X_i) \leq 1 - c$. In words, this additional assumption ensures that there will be observations in both groups having a common value of observed characteristics if the sample size is large enough. The function $p(X_i) = \mathbb{P}(T_i = 1|X_i)$ is known as the *propensity score* and plays an important role in the literature. Finally, it is important to note that for many applications of interest, the model described above employs stronger assumptions than needed. For simplicity, however, the following discussion does not address these distinctions.

This setup naturally motivates matching: observations sharing common (or very similar) values of the observable characteristics $X_i$ are assumed to be free of any selection biases, rendering the statistical inference that uses these observations valid. Of course, matching is not the only way of conducting correct inference in this model. Several parametric, semiparametric, and nonparametric techniques are available, depending on the object of interest and the assumptions imposed. Nonetheless, matching is an attractive procedure because it does not require employing smoothing techniques and appears to be less sensitive to some choices of user-defined tuning parameters.

To describe a matching procedure in detail, consider the special case of matching that uses the Euclidean distance to obtain $M \geq 1$ matches with replacement for the two groups of observations defined by $T_i = 0$ and $T_i = 1$, using as a reservoir of potential matches for each unit $i$ the group opposite to the group this unit belongs to. Then, for unit $i$ the $m$th match, $m = 1, 2, \ldots, M$ is given by the observation having index $j_m(i)$ such that

$$T_{j_m(i)} \neq T_i \text{ and}$$
$$\sum_{j=1}^{n} \mathbb{1}\{T_j \neq T_i\} \mathbb{1}\{\|X_j - X_i\| \leq \|X_{j_m(i)} - X_i\|\} = m.$$

(The function $\mathbb{1}\{\cdot\}$ is the indicator function and $\|\cdot\|$ represents the Euclidean norm.) In words, for the $i$th unit, the $m$th match corresponds to the $m$th nearest neighbor among those observations belonging to the opposite group of unit $i$, as measured by the Euclidean distance between their observable characteristics. For example, if $m = 1$, then $j_1(i)$ corresponds to the unit's index in the opposite group of unit $i$ with the property that $\|X_{j_1(i)}, -X_i\| \leq \|X_j - X_i\|$ for all $j$ such that $T_j \neq T_i$, that is, $X_{j_m(i)}$ is the observation closest to $X_i$ among all the observations in the appropriate group. Similarly, $X_{j_1(i)}, X_{j_2(i)}, \ldots, X_{j_M(i)}$ are the second closest, third closest, and so forth, observations to $X_i$, among those observations in the appropriate subsample. Notice that to simplify the discussion, this definition assumes existence and uniqueness of an observation with index $j_m(i)$. (It is possible to modify the matching procedure to account for these problems.)

In general, the always observed random vector $X_i$ may include both discrete and continuous random variables. When the distribution of (a subvector of) $X_i$ is discrete, the matching procedure may be done exactly in large samples, leading to so-called *exact matching*. However, for those components of $X_i$ that are continuously distributed, matching cannot be done exactly, and therefore in any given sample there will be a discrepancy in terms of observable characteristics, sometimes called the *matching discrepancy*. This discrepancy generates a bias that may affect inference even asymptotically.

The $M$ matches for unit $i$ are given by the observations with indexes $J_M(i) = \{j_1(i), \ldots, j_M(i)\}$, that

is, $\left(Y_{j_1(i)}, X_{j_1(i)}\right), \ldots, \left(Y_{j_M(i)}, X_{j_M(i)}\right)$. This procedure is repeated for the appropriate subsample of units to obtain the final matched sample. Once the matched sample is available, the statistical procedure of interest may be computed. To this end, the first step is to "recover" those counterfactual variables not observed for each unit, which in the context of matching is done by imputation. For example, first define

$$\hat{Y}_{0i} = \begin{cases} Y_i & \text{if} \quad T_i = 0 \\ \dfrac{1}{M} \displaystyle\sum_{j \in J_M(i)} Y_j & \text{if} \quad T_i = 1 \end{cases} \text{ and}$$

$$\hat{Y}_{1i} = \begin{cases} \dfrac{1}{M} \displaystyle\sum_{j \in J_M(i)} Y_j & \text{if} \quad T_i = 0 \\ Y_i & \text{if} \quad T_i = 1 \end{cases}$$

that is, for each unit the unobserved counterfactual variable is imputed using the average of its $M$ matches. Then simple matching estimators are easy to construct: A matching estimator for $\mu_1 = \mathbb{E}[Y_{1i}]$, the mean of $Y_{1i}$ is given by $\hat{\mu}_1 = \frac{1}{n} \sum_{i=1}^{n} \hat{Y}_{1i}$, while a matching estimator for $\tau = \mu_1 - \mu_0 = \mathbb{E}[Y_{1i}] - \mathbb{E}[Y_{0i}]$, the difference in means between both groups, is given by $\hat{\tau} = \hat{\mu}_1 - \hat{\mu}_0$, where $\hat{\mu}_0 = \frac{1}{n} \sum_{i=1}^{n} \hat{Y}_{0i}$. The latter estimand is called the *average treatment effect* in the literature of program evaluation and has received special attention in the theoretical literature of matching estimation.

Matching may also be carried out using estimated rather than observed random variables. A classical example is the so-called *propensity score matching,* which constructs a matched sample using the estimated propensity score (rather than the observed $X_i$) to measure the proximity between observations. Furthermore, matching may also be used to estimate other population parameters of interest, such as quantiles or dispersion measures, in a conceptually similar way. Intuitively, in all cases a matching estimator imputes values for otherwise unobserved random variables using the matched sample. This imputation procedure coincides with an $M$ nearest neighbor ($M - NN$) nonparametric regression estimator.

The implementation of matching is based on several user-defined options (metric, number of matches, etc.), and therefore numerous variants of this procedure may be considered. In all cases, a fast and reliable algorithm is needed to construct a matched sample. Among the available implementations, the so-called *genetic matching,* which uses evolutionary genetic algorithms to construct the matched sample, appears to work well with moderate sample sizes. This implementation allows for a generalized notion of distance (a reweighted Euclidean norm that includes the Mahalanobis metric as a particular case) and an arbitrary number of matches with and without replacement.

There exist several generalizations of the basic matching procedure described above, a particularly important one being the so-called *optimal full matching*. This procedure generalizes the idea of pair or M matching by constructing multiple submatched samples that may include more than one observation from each group. This procedure encompasses the simple matching procedures previously discussed and enjoys certain demonstrable optimality properties.

## Statistical Inference

In recent years, there have been important theoretical developments in statistics and econometrics concerning matching estimators for average treatment effects under the conditional independence assumption. These results establish the validity and lack of validity of commonly used statistical inference procedures involving simple matching estimators.

Despite the fact that in some cases, and under somewhat restrictive assumptions, exact (finite sample) statistical inference results for matching estimators exist, the most important theoretical developments currently available have been derived for large samples and under mild, standard assumptions. Naturally, these asymptotic results have the advantage of being invariant to particular distributional assumptions and the disadvantage of being valid only for large enough samples.

First, despite the relative complexity of matching estimators, it has been established that these estimators for averages with and without replacement enjoy root-$n$ consistency and asymptotic normality under reasonable assumptions. In other words, the estimators described in the previous section (as well as other variants of them) achieve the parametric rate of convergence having a Gaussian limiting distribution after appropriate centering and rescaling. It is important to note that the necessary conditions for this result to hold include the restriction that at most one dimension of the observed characteristics is continuously distributed, regardless of how many

discrete covariates are included in the vector of observed characteristics used by the matching procedure. Intuitively, this restriction arises as a consequence of the bias introduced by the matching discrepancy for continuously distributed observed characteristics, which turns out not to vanish even asymptotically when more than one continuous covariate are included. This problem may be fixed at the expense of introducing further bias reduction techniques that involve nonparametric smoothing procedures, making the "bias corrected" matching estimator somehow less appealing.

Second, regarding the (asymptotic) precision of matching estimators for averages, it has been shown that these estimators do not achieve the minimum possible variance, that is, these estimators are inefficient when compared with other available procedures. However, this efficiency loss is relatively small and decreases fast with the number of matches to be found for each observation.

Finally, in terms of uncertainty estimates of matching estimators for averages, two important results are available. First, it has been shown that the classical *bootstrap* procedure would provide an inconsistent estimate of the standard errors of the matching estimators. For this reason, other resampling techniques must be used, such as *m out of n bootstrap* or *subsampling,* which do deliver consistent standard error estimates under mild regularity conditions. Second, as an alternative, it is possible to construct a consistent estimator of the standard errors that does not require explicit estimation of nonparametric parameters. This estimator uses the matched sample to construct a consistent estimator of the asymptotic (two-piece) variance of the matching estimator.

In sum, the main theoretical results available justify asymptotically the use of classical inference procedures based on the normal distribution, provided the standard errors are estimated appropriately. Computer programs implementing matching, which also compute matching estimators as well as other statistical procedures based on a matched sample, are available in commonly used statistical computing software such as MATLAB, R, and Stata.

*Matias D. Cattaneo*

***See also*** Observational Research; Propensity Score Analysis; Selection

**Further Readings**

Abadie, A., & Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica, 74,* 235–267.

Abadie, A., & Imbens, G. W. (2008). On the failure of the bootstrap for matching estimators. *Econometrica, 76,* 1537–1557.

Abadie, A., & Imbens, G. W. (2009, February). *A martingale representation for matching estimators.* National Bureau of Economic Research, working paper 14756.

Diamond, A., & Sekhon, J. S. (2008, December). *Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies.* Retrieved February 12, 2010, from http://sekhon.berkeley.edu

Hansen, B. B., & Klopfer, S. O. (2006). Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics, 15,* 609–627.

Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature, 47,* 5–86.

Rosenbaum, P. (2002). *Observational studies.* New York: Springer.

# Matrix Algebra

James Joseph Sylvester developed the modern concept of matrices in the 19th century. For him a matrix was an array of numbers. He worked with systems of linear equations; matrices provided a convenient way of working with their *coefficients,* and matrix algebra was to generalize number operations to matrices. Nowadays, matrix algebra is used in all branches of mathematics and the sciences and constitutes the basis of most statistical procedures.

## Matrices: Definition

A matrix is a set of numbers arranged in a table. For example, Toto, Marius, and Olivette are looking at their possessions, and they are counting how many balls, cars, coins, and novels they each possess. Toto has 2 balls, 5 cars, 10 coins, and 20 novels. Marius has 1, 2, 3, and 4, and Olivette has 6, 1, 3, and 10. These data can be displayed in a table in which each row represents a person and each column a possession:

| *Person* | *Balls* | *Cars* | *Coins* | *Novels* |
|---|---|---|---|---|
| Toto | 2 | 5 | 10 | 20 |
| Marius | 1 | 2 | 3 | 4 |
| Olivette | 6 | 1 | 3 | 10 |

We can also say that these data are described by the matrix denoted **A** equal to

$$\mathbf{A} = \begin{bmatrix} 2 & 5 & 10 & 20 \\ 1 & 2 & 3 & 4 \\ 6 & 1 & 3 & 10 \end{bmatrix}. \quad (1)$$

Matrices are denoted by boldface uppercase letters.

To identify a specific element of a matrix, we use its row and column numbers. For example, the cell defined by row 3 and column 1 contains the value 6. We write that $a_{3,1} = 6$. With this notation, elements of a matrix are denoted with the same letter as the matrix but written in lowercase italic. The first subscript always gives the row number of the element (i.e., 3), and second subscript always gives its column number (i.e., 1).

A generic element of a matrix is identified with indices such as $i$ and $j$. So, $a_{i,j}$ is the element at the $i$th row and $j$th column of **A**. The *total* number of rows and columns is denoted with the same letters as the indices but in uppercase letters. The matrix **A** has $I$ rows (here $I = 3$) and $J$ columns (here $J = 4$), and it is made of $I \times J$ elements $a_{i,j}$ (here $3 \times 4 = 12$). The term *dimensions* is often used to refer to the number of rows and columns, so **A** has dimensions $I$ by $J$.

As a shortcut, a matrix can be represented by its generic element written in brackets. So **A** with $I$ rows and $J$ columns is denoted

$$\mathbf{A} = [a_{i,j}] = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,j} & \cdots & a_{1,J} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,j} & \cdots & a_{2,J} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{i,1} & a_{i,2} & \cdots & a_{i,j} & \cdots & a_{i,J} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{I,1} & a_{I,2} & \cdots & a_{I,j} & \cdots & a_{I,J} \end{bmatrix} \quad (2)$$

For either convenience or clarity, the number of rows and columns can also be indicated as subscripts below the matrix name:

$$\mathbf{A} = \underset{I \times J}{\mathbf{A}} = [a_{i,j}]. \quad (3)$$

### *Vectors*

A matrix with one column is called a *column vector* or simply a vector. Vectors are denoted with bold lowercase letters. For example, the first column of matrix **A** (of Equation 1) is a column vector that stores the number of balls of Toto, Marius, and Olivette. We can call it **b** (for balls), and so

$$\mathbf{b} = \begin{bmatrix} 2 \\ 1 \\ 6 \end{bmatrix}. \quad (4)$$

Vectors are the building blocks of matrices. For example, **A** (of Equation 1) is made of four column vectors, which represent the number of balls, cars, coins, and novels, respectively.

### *Norm of a Vector*

We can associate to a vector a quantity, related to its variance and standard deviation, called the *norm* or *length*. The norm of a vector is the square root of the sum of squares of the elements. It is denoted by putting the name of the vector between a set of double bars (||). For example, for

$$\mathbf{b} = \begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix}, \quad (5)$$

we find

$$||\mathbf{x}|| = \sqrt{2^2 + 1^2 + 2^2} = \sqrt{4 + 1 + 4} = \sqrt{9} = 3 \quad (6)$$

### *Normalization of a Vector*

A vector is normalized when its norm is equal to 1. To normalize a vector, we divide each of its elements by its norm. For example, vector **x** from Equation 5 is transformed into the normalized $\overline{\mathbf{x}}$ as

$$\overline{\mathbf{x}} = \frac{\mathbf{x}}{||\mathbf{x}||} = \begin{bmatrix} \frac{2}{3} \\ \frac{1}{3} \\ \frac{2}{3} \end{bmatrix}. \tag{7}$$

## Operations for Matrices

### *Transposition*

If we exchange the roles of the rows and the columns of a matrix, we *transpose* it. This operation is called the *transposition,* and the new matrix is called a *transposed* matrix. The $\mathbf{A}$ matrix transposed is denoted $\mathbf{A}^{\mathrm{T}}$. For example:

$$\text{if } \mathbf{A} = \underset{3 \times 4}{\mathbf{A}} = \begin{bmatrix} 2 & 5 & 10 & 20 \\ 1 & 2 & 3 & 4 \\ 6 & 1 & 3 & 10 \end{bmatrix}, \text{then}$$

$$\mathbf{A}^{\mathrm{T}} = \underset{4 \times 3}{\mathbf{A}^{\mathrm{T}}} = \begin{bmatrix} 2 & 1 & 6 \\ 5 & 2 & 1 \\ 10 & 3 & 3 \\ 20 & 4 & 10 \end{bmatrix}. \tag{8}$$

### *Addition of Matrices*

When two matrices have the same dimensions, we compute their sum by adding the corresponding elements. For example, with

$$\mathbf{A} = \begin{bmatrix} 2 & 5 & 10 & 20 \\ 1 & 2 & 3 & 4 \\ 6 & 1 & 3 & 10 \end{bmatrix} \text{ and}$$
$$\mathbf{B} = \begin{bmatrix} 3 & 4 & 5 & 6 \\ 2 & 4 & 6 & 8 \\ 1 & 2 & 3 & 5 \end{bmatrix}, \tag{9}$$

we find

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} 2+3 & 5+4 & 10+5 & 20+6 \\ 1+2 & 2+4 & 3+6 & 4+8 \\ 6+1 & 1+2 & 3+3 & 10+5 \end{bmatrix}$$
$$= \begin{bmatrix} 5 & 9 & 15 & 26 \\ 3 & 6 & 9 & 12 \\ 7 & 3 & 6 & 15 \end{bmatrix}. \tag{10}$$

In general

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} a_{1,1}+b_{1,1} & a_{1,2}+b_{1,2} & \cdots & a_{1,j}+b_{1,j} & \cdots & a_{1,J}+b_{1,J} \\ a_{2,1}+b_{2,1} & a_{2,2}+b_{2,2} & \cdots & a_{2,j}+b_{2,j} & \cdots & a_{2,J}+b_{2,J} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{i,1}+b_{i,1} & a_{i,2}+b_{i,2} & \cdots & a_{i,j}+b_{i,j} & \cdots & a_{i,J}+b_{i,J} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{I,1}+b_{I,1} & a_{I,2}+b_{I,2} & \cdots & a_{I,j}+b_{I,j} & \cdots & a_{I,J}+b_{I,J} \end{bmatrix}. \tag{11}$$

Matrix addition behaves very much like usual addition. Specifically, matrix addition is *commutative* (i.e., $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$) and *associative* (i.e., $\mathbf{A} + [\mathbf{B} + \mathbf{C}] = [\mathbf{A} + \mathbf{B}] + \mathbf{C}$).

### *Multiplication of a Matrix by a Scalar*

To differentiate matrices from the usual numbers, we call the latter *scalar numbers* or simply *scalars*. To multiply a matrix by a scalar, multiply each element of the matrix by this scalar. For example:

$$10 \times \mathbf{B} = 10 \times \begin{bmatrix} 3 & 4 & 5 & 6 \\ 2 & 4 & 6 & 8 \\ 1 & 2 & 3 & 5 \end{bmatrix}$$
$$= \begin{bmatrix} 30 & 40 & 50 & 60 \\ 20 & 40 & 60 & 80 \\ 10 & 20 & 30 & 50 \end{bmatrix}. \tag{12}$$

### *Multiplication: Product or Products?*

There are *several* ways of generalizing the concept of *product* to matrices. We will look at the most frequently used of these matrix products. Each of these products will behave like the product between scalars when the matrices have dimensions $1 \times 1$.

### *Hadamard Product*

When generalizing product to matrices, the first approach is to multiply the corresponding elements of the two matrices that we want to

multiply. This is called the *Hadamard product,* denoted by $\odot$. The Hadamard product exists only for matrices with the same dimensions. Formally, it is defined as shown below, in matrix 13:

$$\begin{aligned}\mathbf{A} \odot \mathbf{B} &= [a_{i,j} \times b_{i,j}] \\ &= \begin{bmatrix} a_{1,1} \times b_{1,1} & a_{1,2} \times b_{1,2} & \cdots & a_{1,j} \times b_{1,j} & \cdots & a_{1,J} \times b_{1,J} \\ a_{2,1} \times b_{2,1} & a_{2,2} \times b_{2,2} & \cdots & a_{2,j} \times b_{2,j} & \cdots & a_{2,J} \times b_{2,J} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{i,1} \times b_{i,1} & a_{i,2} \times b_{i,2} & \cdots & a_{i,j} \times b_{i,j} & \cdots & a_{i,J} \times b_{i,J} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{I,1} \times b_{I,1} & a_{I,2} \times b_{I,2} & \cdots & a_{I,j} \times b_{I,j} & \cdots & a_{I,J} \times b_{I,J} \end{bmatrix}. \end{aligned} \tag{13}$$

For example, with

$$\mathbf{A} = \begin{bmatrix} 2 & 5 & 10 & 20 \\ 1 & 2 & 3 & 4 \\ 6 & 1 & 3 & 10 \end{bmatrix} \text{ and } \mathbf{B} = \begin{bmatrix} 3 & 4 & 5 & 6 \\ 2 & 4 & 6 & 8 \\ 1 & 2 & 3 & 5 \end{bmatrix}, \tag{14}$$

we get

$$\mathbf{A} \odot \mathbf{B} = \begin{bmatrix} 2 \times 3 & 5 \times 4 & 10 \times 5 & 20 \times 6 \\ 1 \times 2 & 2 \times 4 & 3 \times 6 & 4 \times 8 \\ 6 \times 1 & 1 \times 2 & 3 \times 3 & 10 \times 5 \end{bmatrix} = \begin{bmatrix} 6 & 20 & 50 & 120 \\ 2 & 8 & 18 & 32 \\ 6 & 2 & 9 & 50 \end{bmatrix}. \tag{15}$$

### *Standard or Cayley Product*

The Hadamard product is straightforward, but it is *not* the matrix product that is used most often. The most often used product is called the *standard* or *Cayley* product, or simply *the* product (i.e., when the name of the product is not specified, it is the standard product). Its definition comes from the original use of matrices to solve equations. Its definition looks surprising at first because it is defined only when the number of columns of the first matrix is equal to the number of rows of the second matrix. When two matrices can be multiplied together, they are called *conformable*. This product will have the number of rows of the *first* matrix and the number of columns of the *second* matrix.

So, $\mathbf{A}$ with $I$ rows and $J$ columns can be multiplied by $\mathbf{B}$ with $J$ rows and $K$ columns to give $\mathbf{C}$ with $I$ rows and $K$ columns. A convenient way of checking that two matrices are conformable is to write the dimensions of the matrices as subscripts. For example,

$$\underset{I \times J}{\mathbf{A}} \times \underset{J \times K}{\mathbf{B}} = \underset{I \times K}{\mathbf{C}}, \tag{16}$$

or even

$$\underset{I}{\mathbf{A}}\ \underset{J}{\mathbf{B}}\ \underset{K}{} = \underset{I \times K}{\mathbf{C}}. \tag{17}$$

An element $c_{i,k}$ of the matrix $\mathbf{C}$ is computed as

$$c_{i,k} = \sum_{j=1}^{J} a_{i,j} \times b_{j,k}. \tag{18}$$

So $c_{i,k}$ is the sum of $J$ terms, each term being the product of the corresponding element of the $i$th row of $\mathbf{A}$ with the $k$th column of $\mathbf{B}$.

For example, let

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 5 \end{bmatrix} \text{ and } \mathbf{B} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}. \tag{19}$$

The product of these matrices is denoted $\mathbf{C} = \mathbf{A} \times \mathbf{B} = \mathbf{AB}$ (the $\times$ sign can be omitted when

the context is clear). To compute $c_{2,1}$ we add three terms: (1) the product of the first element of the second row of **A** (i.e., 4) with the first element of the first column of **B** (i.e., 1); (2) the product of the second element of the second row of **A** (i.e., 5) with the second element of the first column of **B** (i.e., 3); and (3) the product of the third element of the second row of **A** (i.e., 5) with the third element of the first column of **B** (i.e., 5). Formally, the term $c_{2,1}$ is obtained as

$$\begin{aligned} c_{2,1} &= \sum_{j=1}^{J=3} a_{2,j} \times b_{j,1} \\ &= (a_{2,1}) \times (b_{1,1}) + (a_{2,2} \times b_{2,1}) + (a_{2,3} \times b_{3,1}) \\ &= (4 \times 1) + (5 \times 3) + (6 \times 5) \\ &= 49. \end{aligned} \tag{20}$$

Matrix **C** is obtained as

$$\begin{aligned} \mathbf{AB} &= \\ \mathbf{C} &= [c_{i,k}] \\ &= \sum_{j=1}^{J=3} a_{i,j} \times b_{j,k} \\ &= \begin{bmatrix} 1\times1+2\times3+3\times5 & 1\times2+2\times4+3\times6 \\ 4\times1+5\times3+6\times5 & 4\times2+5\times4+6\times6 \end{bmatrix} \\ &= \begin{bmatrix} 22 & 28 \\ 49 & 64 \end{bmatrix}. \end{aligned} \tag{21}$$

### *Properties of the Product*

Like the product between scalars, the product between matrices is *associative,* and *distributive* relative to addition. Specifically, for any set of three conformable matrices **A**, **B**, and **C**,

$$(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC}) = \mathbf{ABC} \quad \text{associativity} \tag{22}$$

$$\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC} \qquad \text{distributivity.} \tag{23}$$

The matrix products **AB** and **BA** do not always exist, but when they do, these products are *not,* in general, *commutative:*

$$\mathbf{AB} \neq \mathbf{BA}. \tag{24}$$

For example, with

$$\mathbf{A} = \begin{bmatrix} 2 & 1 \\ -2 & -1 \end{bmatrix} \text{ and } \mathbf{B} = \begin{bmatrix} 1 & -1 \\ -2 & 2 \end{bmatrix} \tag{25}$$

we get:

$$\mathbf{AB} = \begin{bmatrix} 2 & 1 \\ -2 & -1 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ -2 & 2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}. \tag{26}$$

But

$$\begin{aligned} \mathbf{BA} &= \begin{bmatrix} 1 & -1 \\ -2 & 2 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ -2 & -1 \end{bmatrix} \\ &= \begin{bmatrix} 4 & 2 \\ -8 & -4 \end{bmatrix}. \end{aligned} \tag{27}$$

Incidently, we can combine transposition and product and get the following equation:

$$(\mathbf{AB})^{\mathrm{T}} = \mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}. \tag{28}$$

### *Exotic Product: Kronecker*

Another product is the *Kronecker* product, also called the *direct, tensor,* or *Zehfuss* product. It is denoted $\otimes$ and is defined for all matrices. Specifically, with two matrices $\mathbf{A} = a_{i,j}$ (with dimensions $I$ by $J$) and **B** (with dimensions $K$ and $L$), the Kronecker product gives a matrix **C** (with dimensions $(I \times K)$ by $(J \times L)$ defined as

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{1,1}\mathbf{B} & a_{1,2}\mathbf{B} & \cdots & a_{1,j}\mathbf{B} & \cdots & a_{1,J}\mathbf{B} \\ a_{2,1}\mathbf{B} & a_{2,2}\mathbf{B} & \cdots & a_{2,j}\mathbf{B} & \cdots & a_{2,J}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{i,1}\mathbf{B} & a_{i,2}\mathbf{B} & \cdots & a_{i,j}\mathbf{B} & \cdots & a_{i,J}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{I,1}\mathbf{B} & a_{I,2}\mathbf{B} & \cdots & a_{I,j}\mathbf{B} & \cdots & a_{I,J}\mathbf{B} \end{bmatrix}. \tag{29}$$

For example, with

$$\mathbf{A} = [\,1 \quad 2 \quad 3\,] \text{ and } \mathbf{B} = \begin{bmatrix} 6 & 7 \\ 8 & 9 \end{bmatrix} \tag{30}$$

we get

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} 1\times 6 & 1\times 7 & 2\times 6 & 2\times 7 & 3\times 6 & 3\times 7 \\ 1\times 8 & 1\times 9 & 2\times 8 & 2\times 9 & 3\times 8 & 3\times 9 \end{bmatrix} = \begin{bmatrix} 6 & 7 & 12 & 14 & 18 & 21 \\ 8 & 9 & 16 & 18 & 24 & 27 \end{bmatrix}. \tag{31}$$

The Kronecker product is used to write design matrices. It is an essential tool for the derivation of expected values and sampling distributions.

## Special Matrices

Certain special matrices have specific names.

### *Square and Rectangular Matrices*

A matrix with the same number of rows and columns is a *square matrix*. By contrast, a matrix with different numbers of rows and columns is a *rectangular matrix*. So

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 5 \\ 7 & 8 & 0 \end{bmatrix} \tag{32}$$

is a square matrix, but

$$\mathbf{B} = \begin{bmatrix} 1 & 2 \\ 4 & 5 \\ 7 & 8 \end{bmatrix} \tag{33}$$

is a rectangular matrix.

### *Symmetric Matrix*

A square matrix $\mathbf{A}$ with $a_{i,j} = a_{j,i}$ is *symmetric*. So

$$\mathbf{A} = \begin{bmatrix} 10 & 2 & 3 \\ 2 & 20 & 5 \\ 3 & 5 & 30 \end{bmatrix} \tag{34}$$

is symmetric, but

$$\mathbf{A} = \begin{bmatrix} 12 & 2 & 3 \\ 4 & 20 & 5 \\ 7 & 8 & 30 \end{bmatrix} \tag{35}$$

is not.

Note that for a symmetric matrix,

$$\mathbf{A} = \mathbf{A}^{\mathrm{T}}. \tag{36}$$

A common mistake is to assume that the standard product of two symmetric matrices is commutative. But this is not true, as shown by the following example. With

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 1 & 4 \\ 3 & 4 & 1 \end{bmatrix} \text{ and } \mathbf{B} = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 1 & 3 \\ 2 & 3 & 1 \end{bmatrix} \tag{37}$$

we get

$$\mathbf{AB} = \begin{bmatrix} 9 & 12 & 11 \\ 11 & 15 & 11 \\ 9 & 10 & 19 \end{bmatrix}, \text{ but } \mathbf{BA} = \begin{bmatrix} 9 & 11 & 9 \\ 12 & 15 & 10 \\ 11 & 11 & 19 \end{bmatrix}. \tag{38}$$

Note, however, that combining Equations 35 and 43 gives for symmetric matrices $\mathbf{A}$ and $\mathbf{B}$ the following equation:

$$\mathbf{AB} = (\mathbf{BA})^{\mathrm{T}}. \tag{39}$$

### Diagonal Matrix

A square matrix is *diagonal* when all its elements except the ones on the diagonal are zero. Formally, a matrix is diagonal if $a_{i,j} = 0$ when $i \neq j$. Thus

$$\mathbf{A} = \begin{bmatrix} 10 & 0 & 0 \\ 0 & 20 & 0 \\ 0 & 0 & 30 \end{bmatrix} \text{ is diagonal.} \tag{40}$$

Because only the diagonal elements matter for a diagonal matrix, we can specify only these diagonal elements. This is done with the following notation:

$$\mathbf{A} = \mathrm{diag}\{[a_{1,1}, \ldots, a_{i,i}, \ldots, a_{I,I}]\} = \mathrm{diag}\{[a_{i,i}]\}. \tag{41}$$

For example, the previous matrix can be rewritten as:

$$\mathbf{A} = \begin{bmatrix} 10 & 0 & 0 \\ 0 & 20 & 0 \\ 0 & 0 & 30 \end{bmatrix} = \text{diag}\{[10, 20, 30]\}. \quad (42)$$

The operator *diag* can also be used to isolate the diagonal of any square matrix. For example, with

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} \quad (43)$$

we get

$$\text{diag}\{\mathbf{A}\} = \text{diag}\left\{\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}\right\} = \begin{bmatrix} 1 \\ 5 \\ 9 \end{bmatrix}. \quad (44)$$

Note, incidentally, that

$$\text{diag}\{\text{diag}\{\mathbf{A}\}\} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 9 \end{bmatrix}. \quad (45)$$

### *Multiplication by a Diagonal Matrix*

Diagonal matrices are often used to multiply by a scalar all the elements of a given row or column. Specifically, when we premultiply a matrix by a diagonal matrix, the elements of the row of the second matrix are multiplied by the corresponding diagonal element. Likewise, when we postmultiply a matrix by a diagonal matrix, the elements of the column of the first matrix are multiplied by the corresponding diagonal element. For example, with:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} 2 & 0 \\ 0 & 5 \end{bmatrix}$$
$$\mathbf{C} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 6 \end{bmatrix}, \quad (46)$$

we get

$$\mathbf{BA} = \begin{bmatrix} 2 & 0 \\ 0 & 5 \end{bmatrix} \times \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} = \begin{bmatrix} 2 & 4 & 6 \\ 20 & 25 & 30 \end{bmatrix} \quad (47)$$

and

$$\mathbf{AC} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \times \begin{bmatrix} 2 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 6 \end{bmatrix} = \begin{bmatrix} 2 & 8 & 18 \\ 8 & 20 & 36 \end{bmatrix} \quad (48)$$

and also

$$\mathbf{BAC} = \begin{bmatrix} 2 & 0 \\ 0 & 5 \end{bmatrix} \times \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \times \begin{bmatrix} 2 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 6 \end{bmatrix} = \begin{bmatrix} 4 & 16 & 36 \\ 40 & 100 & 180 \end{bmatrix}. \quad (49)$$

### *Identity Matrix*

A diagonal matrix whose diagonal elements are all equal to 1 is called an *identity* matrix and is denoted **I**. If we need to specify its dimensions, we use subscripts such as

$$\underset{3 \times 3}{\mathbf{I}} = \mathbf{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \text{ (this is a } 3 \times 3 \text{ identity matrix).} \quad (50)$$

The identity matrix is the neutral element for the standard product. So

$$\mathbf{I} \times \mathbf{A} = \mathbf{A} \times \mathbf{I} = \mathbf{A} \quad (51)$$

for any matrix **A** conformable with **I**. For example:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 5 \\ 7 & 8 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 5 \\ 7 & 8 & 0 \end{bmatrix} \times \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 5 \\ 7 & 8 & 0 \end{bmatrix}. \quad (52)$$

*Matrix Full of Ones*

A matrix whose elements are all equal to 1 is denoted by $\mathbf{1}$ or, when we need to specify its dimensions, by $\underset{I \times J}{\mathbf{1}}$. These matrices are neutral elements for the Hadamard product. So

$$\underset{2 \times 3}{\mathbf{A}} \odot \underset{2 \times 3}{\mathbf{1}} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \odot \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \quad (53)$$

$$= \begin{bmatrix} 1 \times 1 & 2 \times 1 & 3 \times 1 \\ 4 \times 1 & 5 \times 1 & 6 \times 1 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}. \quad (54)$$

The matrices can also be used to compute sums of rows or columns:

$$\begin{bmatrix} 1 & 2 & 3 \end{bmatrix} \times \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = (1 \times 1) + (2 \times 1) + (3 \times 1) = 1 + 2 + 3 = 6, \quad (55)$$

or also

$$\begin{bmatrix} 1 & 1 \end{bmatrix} \times \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} = \begin{bmatrix} 5 & 7 & 9 \end{bmatrix}. \quad (56)$$

*Matrix Full of Zeros*

A matrix whose elements are all equal to 0 is the *null* or *zero* matrix. It is denoted by 0 or, when we need to specify its dimensions, by $\underset{I \times J}{\mathbf{0}}$. Null matrices are neutral elements for addition:

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} + \underset{2 \times 2}{\mathbf{0}} = \begin{bmatrix} 1+0 & 2+0 \\ 3+0 & 4+0 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}. \quad (57)$$

They are also null elements for the Hadamard product:

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \odot \underset{2 \times 2}{\mathbf{0}} = \begin{bmatrix} 1 \times 0 & 2 \times 0 \\ 3 \times 0 & 4 \times 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} = \underset{2 \times 2}{\mathbf{0}} \quad (58)$$

and for the standard product:

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \times \underset{2 \times 2}{\mathbf{0}} = \begin{bmatrix} 1 \times 0 + 2 \times 0 & 1 \times 0 + 2 \times 0 \\ 3 \times 0 + 4 \times 0 & 3 \times 0 + 4 \times 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} = \underset{2 \times 2}{\mathbf{0}}. \quad (59)$$

*Triangular Matrix*

A matrix is lower triangular when $a_{i,j} = 0$ for $i < j$. A matrix is upper triangular when $a_{i,j} = 0$ for $i > j$. For example,

$$\mathbf{A} = \begin{bmatrix} 10 & 0 & 0 \\ 2 & 20 & 0 \\ 3 & 5 & 30 \end{bmatrix} \text{ is lower triangular,} \quad (60)$$

and

$$\mathbf{B} = \begin{bmatrix} 12 & 2 & 3 \\ 0 & 20 & 5 \\ 0 & 0 & 30 \end{bmatrix} \text{ is upper triangular.} \quad (61)$$

*Cross-Product Matrix*

A cross-product matrix is obtained by multiplication of a matrix by its transpose. Therefore a cross-product matrix is square and symmetric. For example, the matrix:

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 2 & 4 \\ 3 & 4 \end{bmatrix} \quad (62)$$

premultiplied by its transpose

$$\mathbf{A}^{\mathrm{T}} = \begin{bmatrix} 1 & 2 & 3 \\ 1 & 4 & 4 \end{bmatrix} \quad (63)$$

gives the cross-product matrix

$$\mathbf{A}^{\mathrm{T}}\mathbf{A} = \begin{bmatrix} 1 \times 1 + 2 \times 2 + 3 \times 3 & 1 \times 1 + 2 \times 4 + 3 \times 4 \\ 1 \times 1 + 4 \times 2 + 4 \times 3 & 1 \times 1 + 4 \times 4 + 4 \times 4 \end{bmatrix} = \begin{bmatrix} 14 & 21 \\ 14 & 33 \end{bmatrix}. \quad (64)$$

### *A Particular Case of Cross-Product Matrix: Variance–Covariance*

A particular case of cross-product matrices is correlation or covariance matrices. A variance–covariance matrix is obtained from a data matrix by three steps: (1) subtract the mean of each column from each element of this column (this is *centering*), (2) compute the cross-product matrix from the centered matrix, and (3) divide each element of the cross-product matrix by the number of rows of the data matrix. For example, if we take the $I = 3$ by $J = 2$ matrix $\mathbf{A}$,

$$\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 5 & 10 \\ 8 & 10 \end{bmatrix}, \tag{65}$$

we obtain the means of each column as

$$\begin{aligned} \mathbf{m} &= \frac{1}{I} \times \underset{1 \times I}{\mathbf{1}} \times \underset{I \times J}{\mathbf{A}} \\ &= \frac{1}{3} \times [1 \quad 1 \quad 1] \times \begin{bmatrix} 2 & 1 \\ 5 & 10 \\ 8 & 10 \end{bmatrix} = [5 \quad 7]. \end{aligned} \tag{66}$$

To center the matrix, we subtract the mean of each column from all its elements. This centered matrix gives the deviations of each element from the mean of its column. Centering is performed as

$$\mathbf{D} = \mathbf{A} - \underset{J \times 1}{\mathbf{1}} \times \mathbf{m} = \begin{bmatrix} 2 & 1 \\ 5 & 10 \\ 8 & 10 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \times [5 \quad 7] \tag{67}$$

$$= \begin{bmatrix} 2 & 1 \\ 5 & 10 \\ 8 & 10 \end{bmatrix} - \begin{bmatrix} 5 & 7 \\ 5 & 7 \\ 5 & 7 \end{bmatrix} = \begin{bmatrix} -3 & -6 \\ 0 & 3 \\ 3 & 3 \end{bmatrix}. \tag{68}$$

We denote as $\mathbf{S}$ the variance–covariance matrix derived from $\mathbf{A}$. It is computed as

$$\begin{aligned} \mathbf{S} = \frac{1}{I}\mathbf{D}^{\mathrm{T}}\mathbf{D} &= \frac{1}{3}\begin{bmatrix} -3 & 0 & 3 \\ -6 & 3 & 3 \end{bmatrix} \times \begin{bmatrix} -3 & -6 \\ 0 & 3 \\ 3 & 3 \end{bmatrix} \\ &= \frac{1}{3} \times \begin{bmatrix} 18 & 27 \\ 27 & 54 \end{bmatrix} = \begin{bmatrix} 6 & 9 \\ 9 & 18 \end{bmatrix}. \end{aligned} \tag{69}$$

(Variances are on the diagonal; covariances are off-diagonal.)

## The Inverse of a Square Matrix

An operation similar to division exists, but only for (some) square matrices. This operation uses the notion of *inverse* operation and defines the *inverse* of a matrix. The inverse is defined by analogy with the scalar number case, for which division actually corresponds to multiplication by the inverse, namely,

$$\frac{a}{b} = a \times b^{-1} \text{ with } b \times b^{-1} = 1. \tag{70}$$

The inverse of a square matrix $\mathbf{A}$ is denoted $\mathbf{A}^{-1}$. It has the following property:

$$\mathbf{A} \times \mathbf{A}^{-1} = \mathbf{A}^{-1} \times \mathbf{A} = \mathbf{I}. \tag{71}$$

The definition of the inverse of a matrix is simple, but its computation is complicated and is best left to computers.

For example, for

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{72}$$

The inverse is:

$$\mathbf{A}^{-1} = \begin{bmatrix} 1 & -2 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{73}$$

All square matrices do not necessarily have an inverse. The inverse of a matrix does not exist if the rows (and the columns) of this matrix are linearly dependent. For example,

$$\mathbf{A} = \begin{bmatrix} 3 & 4 & 2 \\ 1 & 0 & 2 \\ 2 & 1 & 3 \end{bmatrix} \tag{74}$$

does not have an inverse since the second column is a linear combination of the two other columns:

$$\begin{bmatrix} 4 \\ 0 \\ 1 \end{bmatrix} = 2 \times \begin{bmatrix} 3 \\ 1 \\ 2 \end{bmatrix} - \begin{bmatrix} 2 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 6 \\ 2 \\ 4 \end{bmatrix} - \begin{bmatrix} 2 \\ 2 \\ 3 \end{bmatrix}. \tag{75}$$

A matrix without an inverse is *singular*. When $\mathbf{A}^{-1}$ exists, it is unique.

Inverse matrices are used for solving linear equations and least square problems in multiple regression analysis or analysis of variance.

### *Inverse of a Diagonal Matrix*

The inverse of a diagonal matrix is easy to compute: The inverse of

$$\mathbf{A} = \text{diag}\{a_{i,i}\} \tag{76}$$

is the diagonal matrix

$$\mathbf{A}^{-1} = \text{diag}\left\{a_{i,i}^{-1}\right\} = \text{diag}\{1/a_{i,i}\}. \tag{77}$$

For example,

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & .5 & 0 \\ 0 & 0 & 4 \end{bmatrix} \text{ and } \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & .25 \end{bmatrix} \tag{78}$$

are the inverse of each other.

## The Big Tool: Eigendecomposition

So far, matrix operations are very similar to operations with numbers. The next notion is specific to matrices. This is the idea of decomposing a matrix into simpler matrices. A lot of the power of matrices follows from this. A first decomposition is called the *eigendecomposition*, and it applies only to square matrices. The generalization of the eigendecomposition to rectangular matrices is called the *singular value* decomposition.

*Eigenvectors* and *eigenvalues* are numbers and vectors associated with square matrices. Together they constitute the *eigendecomposition*. Even though the eigendecomposition does not exist for all square matrices, it has a particularly simple expression for a class of matrices often used in multivariate analysis such as correlation, covariance, or cross-product matrices.The eigendecomposition of these matrices is important in statistics because it is used to find the maximum (or minimum) of functions involving these matrices. For example, principal components analysis is obtained from the eigendecomposition of a covariance or correlation matrix and gives the least square estimate of the original data matrix.

### *Notations and Definition*

An eigenvector of matrix $\mathbf{A}$ is a vector $\mathbf{u}$ that satisfies the following equation:

$$\mathbf{A}\mathbf{u} = \lambda\mathbf{u}, \tag{79}$$

where $\lambda$ is a scalar called the *eigenvalue* associated to the *eigenvector*. When rewritten, Equation 79 becomes

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{u} = 0. \tag{80}$$

Therefore $\mathbf{u}$ is eigenvector of $\mathbf{A}$ if the multiplication of $\mathbf{u}$ by $\mathbf{A}$ changes the length of $\mathbf{u}$ but not its orientation. For example,

$$\mathbf{A} = \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \tag{81}$$

has for eigenvectors

$$\mathbf{u}_1 = \begin{bmatrix} 3 \\ 2 \end{bmatrix} \quad \text{with eigenvalue} \quad \lambda_1 = 4 \tag{82}$$

and

$$\mathbf{u}_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix} \quad \text{with eigenvalue} \quad \lambda_2 = -1\,. \tag{83}$$

When $\mathbf{u}_1$ and $\mathbf{u}_2$ are multiplied by $\mathbf{A}$, only their length changes. That is,

$$\mathbf{A}\mathbf{u}_1 = \lambda_1\mathbf{u}_1 = \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix}\begin{bmatrix} 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 12 \\ 8 \end{bmatrix} = 4\begin{bmatrix} 3 \\ 2 \end{bmatrix} \tag{84}$$

and

$$\mathbf{A}\mathbf{u}_2 = \lambda_2\mathbf{u}_2 = \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix}\begin{bmatrix} -1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix} = -1\begin{bmatrix} -1 \\ 1 \end{bmatrix}. \tag{85}$$

This is illustrated in Figure 1.

For convenience, eigenvectors are generally normalized such that

$$\mathbf{u}^{\mathrm{T}}\mathbf{u} = 1. \tag{86}$$

For the previous example, normalizing the eigenvectors gives

$$\mathbf{u}_1 = \begin{bmatrix} .8321 \\ .5547 \end{bmatrix} \text{ and } \mathbf{u}_2 = \begin{bmatrix} -.7071 \\ .7071 \end{bmatrix}. \tag{87}$$

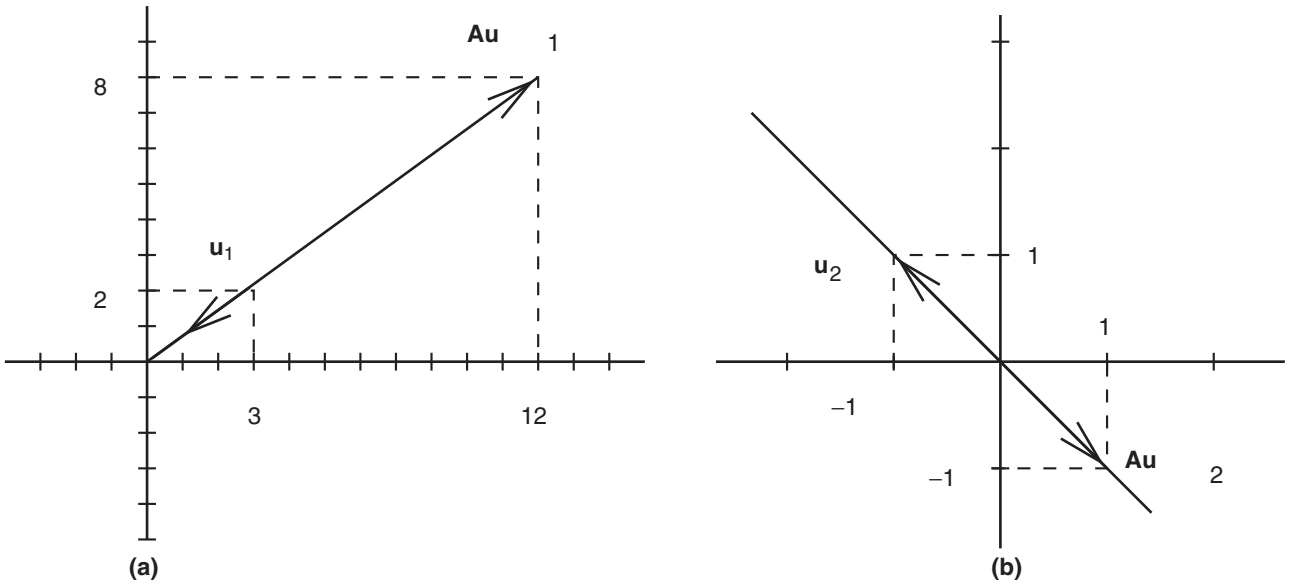



**Figure 1** Two Eigenvectors of a Matrix

We can check that

$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} .8321 \\ .5547 \end{bmatrix} = \begin{bmatrix} 3.3284 \\ 2.2188 \end{bmatrix} = 4 \begin{bmatrix} .8321 \\ .5547 \end{bmatrix} \tag{88}$$

and

$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} -.7071 \\ .7071 \end{bmatrix} = \begin{bmatrix} .7071 \\ -.7071 \end{bmatrix} = -1 \begin{bmatrix} -.7071 \\ .7071 \end{bmatrix}. \tag{89}$$

*Eigenvector and Eigenvalue Matrices*

Traditionally, we store the eigenvectors of $\mathbf{A}$ as the columns of a matrix denoted $\mathbf{U}$. Eigenvalues are stored in a diagonal matrix (denoted $\mathbf{\Lambda}$). Therefore, Equation 79 becomes

$$\mathbf{AU} = \mathbf{U\Lambda}. \tag{90}$$

For example, with $\mathbf{A}$ (from Equation 81), we have

$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \times \begin{bmatrix} 3 & -1 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} 3 & -1 \\ 2 & 1 \end{bmatrix} \times \begin{bmatrix} 4 & 0 \\ 0 & -1 \end{bmatrix}. \tag{91}$$

*Reconstitution of a Matrix*

The eigendecomposition can also be used to build back a matrix from its eigenvectors and eigenvalues. This is shown by rewriting Equation 90 as

$$\mathbf{A} = \mathbf{U\Lambda U}^{-1}. \tag{92}$$

For example, because

$$\mathbf{U}^{-1} = \begin{bmatrix} .2 & .2 \\ -.4 & .6 \end{bmatrix},$$

we obtain

$$\mathbf{A} = \mathbf{U\Lambda U}^{-1} = \begin{bmatrix} 3 & -1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 4 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} .2 & .2 \\ -.4 & .6 \end{bmatrix} = \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix}. \tag{93}$$

*Digression: An Infinity of Eigenvectors for One Eigenvalue*

It is only through a slight abuse of language that we talk about *the* eigenvector associated with *one* eigenvalue. Any scalar multiple of an eigenvector is an eigenvector, so for each eigenvalue, there are

an infinite number of eigenvectors, all proportional to each other. For example,

$$\begin{bmatrix} 1 \\ -1 \end{bmatrix} \tag{94}$$

is an eigenvector of **A**:

$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix}. \tag{95}$$

Therefore,

$$2 \times \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 2 \\ -2 \end{bmatrix} \tag{96}$$

is also an eigenvector of **A**:

$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ -2 \end{bmatrix} = \begin{bmatrix} -2 \\ 2 \end{bmatrix} = -1 \times 2 \begin{bmatrix} 1 \\ -1 \end{bmatrix}. \tag{97}$$

**Positive (Semi)Definite Matrices**

Some matrices, such as $\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$, do not have eigenvalues. Fortunately, the matrices used often in statistics belong to a category called *positive semidefinite*. The eigendecomposition of these matrices always exists and has a particularly convenient form. A matrix is positive semidefinite when it can be obtained as the product of a matrix by its transpose. This implies that a positive semidefinite matrix is always symmetric. So, formally, the matrix **A** is positive semidefinite if it can be obtained as

$$\mathbf{A} = \mathbf{X}\mathbf{X}^{\mathrm{T}} \tag{98}$$

for a certain matrix **X**. Positive semidefinite matrices include correlation, covariance, and cross-product matrices.

The eigenvalues of a positive semidefinite matrix are always positive or null. Its eigenvectors are composed of real values and are pairwise orthogonal when their eigenvalues are different. This implies the following equality:

$$\mathbf{U}^{-1} = \mathbf{U}^{\mathrm{T}}. \tag{99}$$

We can, therefore, express the positive semidefinite matrix **A** as

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}} \tag{100}$$

where $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}$ are the normalized eigenvectors.

For example,

$$\mathbf{A} = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix} \tag{101}$$

can be decomposed as

$$\begin{aligned} \mathbf{A} &= \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}} \\ &= \begin{bmatrix} \sqrt{\frac{1}{2}} & \sqrt{\frac{1}{2}} \\ \sqrt{\frac{1}{2}} & -\sqrt{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} 4 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} \sqrt{\frac{1}{2}} & \sqrt{\frac{1}{2}} \\ \sqrt{\frac{1}{2}} & -\sqrt{\frac{1}{2}} \end{bmatrix} \\ &= \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}, \end{aligned} \tag{102}$$

with

$$\begin{bmatrix} \sqrt{\frac{1}{2}} & \sqrt{\frac{1}{2}} \\ \sqrt{\frac{1}{2}} & -\sqrt{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} \sqrt{\frac{1}{2}} & \sqrt{\frac{1}{2}} \\ \sqrt{\frac{1}{2}} & -\sqrt{\frac{1}{2}} \end{bmatrix} = 1,001. \tag{103}$$

***Diagonalization***

When a matrix is positive semidefinite, we can rewrite Equation 100 as

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}} \quad \Leftrightarrow \quad \Lambda = \mathbf{U}^{\mathrm{T}}\mathbf{A}\mathbf{U}. \tag{104}$$

This shows that we can transform **A** into a *diagonal* matrix. Therefore the eigendecomposition of a positive semidefinite matrix is often called its *diagonalization.*

***Another Definition for Positive Semidefinite Matrices***

A matrix **A** is positive semidefinite if for any nonzero vector **x**, we have

$$\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} \geq 0 \quad \forall \mathbf{x}. \tag{105}$$

When all the eigenvalues of a matrix are positive, the matrix is *positive definite.* In that case, Equation 105 becomes

$$\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} > 0 \quad \forall \mathbf{x}. \tag{106}$$

### *Trace, Determinant, and Rank*

The eigenvalues of a matrix are closely related to three important numbers associated to a square matrix: *trace, determinant,* and *rank*.

#### *Trace*

The trace of **A**, denoted trace{**A**}, is the sum of its diagonal elements. For example, with

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} \tag{107}$$

we obtain

$$\text{trace}\{\mathbf{A}\} = 1 + 5 + 9 = 15. \tag{108}$$

The trace of a matrix is also equal to the sum of its eigenvalues:

$$\text{trace}\{\mathbf{A}\} = \sum_{\ell} \lambda_{\ell} = \text{trace}\{\mathbf{\Lambda}\} \tag{109}$$

with $\mathbf{\Lambda}$ being the matrix of the eigenvalues of **A**. For the previous example, we have

$$\mathbf{\Lambda} = \text{diag}\{16.1168, -1.1168, 0\}. \tag{110}$$

We can verify that

$$\begin{aligned} \text{trace}\{\mathbf{A}\} = \sum_{\ell} \lambda_{\ell} &= 16.1168 \\ &+ (-1.1168) = 15. \end{aligned} \tag{111}$$

#### *Determinant*

The *determinant* is important for finding the solution of systems of linear equations (i.e., the determinant *determines* the existence of a solution). The determinant of a matrix is equal to the product of its eigenvalues. If det{**A**} is the determinant of **A**,

$$\begin{aligned} \det\{\mathbf{A}\} = \prod_{\ell} \lambda_{\ell} \text{ with } \lambda_{\ell} \text{ being the} \\ \ell \text{th eigenvalue of } \mathbf{A}. \end{aligned} \tag{112}$$

For example, the determinant of **A** from Equation 107 is equal to

$$\det\{\mathbf{A}\} = 16.1168 \times -1.1168 \times 0 = 0. \tag{113}$$

#### Rank

Finally, the *rank* of a matrix is the number of nonzero eigenvalues of the matrix. For our example,

$$\text{rank}\{\mathbf{A}\} = 2. \tag{114}$$

The rank of a matrix gives the dimensionality of the Euclidean space that can be used to represent this matrix. Matrices whose rank is equal to their dimensions are *full rank,* and they are invertible. When the rank of a matrix is smaller than its dimensions, the matrix is not invertible and is called *rank-deficient, singular,* or *multicolinear.* For example, matrix **A** from Equation 107 is a $3 \times 3$ square matrix, its rank is equal to 2, and therefore it is rank-deficient and does not have an inverse.

### *Statistical Properties of the Eigendecomposition*

The eigendecomposition is essential in optimization. For example, principal components analysis is a technique used to analyze an $I \times J$ matrix **X** in which the rows are observations and the columns are variables. Principal components analysis finds orthogonal row *factor scores* that "explain" as much of the variance of **X** as possible. They are obtained as

$$\mathbf{F} = \mathbf{XP}, \tag{115}$$

where **F** is the matrix of factor scores and **P** is the matrix of loadings of the variables. These loadings give the coefficients of the linear combination used to compute the factor scores from the variables. In addition to Equation 115, we impose the constraints that

$$\mathbf{F}^{\mathrm{T}}\mathbf{F} = \mathbf{P}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{XP} \tag{116}$$

is a diagonal matrix (i.e., **F** is an orthogonal matrix) and that

$$\mathbf{P}^{\mathrm{T}}\mathbf{P} = \mathbf{I} \tag{117}$$

(i.e., **P** is an orthonormal matrix). The solution is obtained by using Lagrangian multipliers in which the constraint from Equation 117 is expressed as

the multiplication with a diagonal matrix of Lagrangian multipliers denoted $\mathbf{\Lambda}$; in order to give the following expression:

$$\mathbf{\Lambda}\left(\mathbf{P}^{\mathrm{T}}\mathbf{P}-\mathbf{I}\right). \tag{118}$$

This amounts to defining the following equation:

$$\begin{aligned}\mathcal{L} &= \mathbf{F}^{\mathrm{T}}\mathbf{F}-\mathbf{\Lambda}(\mathbf{P}^{\mathrm{T}}\mathbf{P}-\mathbf{I}) \\ &= \mathrm{trace}\{\mathbf{P}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{P}-\mathbf{\Lambda}(\mathbf{P}^{\mathrm{T}}\mathbf{P}-\mathbf{I})\}.\end{aligned} \tag{119}$$

The values of $\mathbf{P}$ that give the maximum values of $\mathcal{L}$ are found by first computing the derivative of $\mathcal{L}$ relative to $\mathbf{P}$,

$$\frac{\partial \mathcal{L}}{\partial \mathbf{P}} = 2\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{P}-2\mathbf{\Lambda}\mathbf{P}, \tag{120}$$

and setting this derivative to zero:

$$\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{P}-\mathbf{\Lambda}\mathbf{P}=0 \Leftrightarrow \mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{P}=\mathbf{\Lambda}\mathbf{P}. \tag{121}$$

Because $\mathbf{\Lambda}$ is diagonal, this is an eigendecomposition problem, $\mathbf{\Lambda}$ is the matrix of eigenvalues of the positive semidefinite matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ ordered from the largest to the smallest, and $\mathbf{P}$ is the matrix of eigenvectors of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$. Finally, the factor matrix is

$$\mathbf{F}=\mathbf{P}\mathbf{\Lambda}^{\frac{1}{2}}. \tag{122}$$

The variance of the factor scores is equal to the eigenvalues:

$$\mathbf{F}^{\mathrm{T}}\mathbf{F}=\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{P}^{\mathrm{T}}\mathbf{P}\mathbf{\Lambda}^{\frac{1}{2}}=\mathbf{\Lambda}. \tag{123}$$

Because the sum of the eigenvalues is equal to the trace of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, the first factor scores "extract" as much of the variance of the original data as possible, the second factor scores extract as much of the variance left unexplained by the first factor as possible, and so on for the remaining factors. The diagonal elements of the matrix $\mathbf{\Lambda}^{\frac{1}{2}}$, which are the standard deviations of the factor scores, are called the *singular values* of $\mathbf{X}$.

## A Tool for Rectangular Matrices: The Singular Value Decomposition

The singular value decomposition (SVD) generalizes the eigendecomposition to rectangular matrices. The eigendecomposition decomposes a matrix into *two* simple matrices, and the SVD decomposes a rectangular matrix into *three* simple matrices: two orthogonal matrices and one diagonal matrix. The SVD uses the eigendecomposition of a positive semidefinite matrix to derive a similar decomposition for rectangular matrices.

### *Definitions and Notations*

The SVD decomposes matrix $\mathbf{A}$ as

$$\mathbf{A}=\mathbf{P}\mathbf{\Delta}\mathbf{Q}^{\mathrm{T}}, \tag{124}$$

where $\mathbf{P}$ is the (normalized) eigenvectors of the matrix $\mathbf{A}\mathbf{A}^{\mathrm{T}}$ (i.e., $\mathbf{P}^{\mathrm{T}}\mathbf{P}=\mathbf{I}$). The columns of $\mathbf{P}$ are called the *left singular vectors* of $\mathbf{A}$. $\mathbf{Q}$ is the (normalized) eigenvectors of the matrix $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ (i.e., $\mathbf{Q}^{\mathrm{T}}\mathbf{Q}=\mathbf{I}$). The columns of $\mathbf{Q}$ are called the *right singular vectors* of $\mathbf{A}$. $\mathbf{\Delta}$ is the diagonal matrix of the *singular values*, $\mathbf{\Delta}=\mathbf{\Lambda}^{\frac{1}{2}}$, with $\mathbf{\Lambda}$ being the diagonal matrix of the eigenvalues of $\mathbf{A}\mathbf{A}^{\mathrm{T}}$ and $\mathbf{A}^{\mathrm{T}}\mathbf{A}$.

The SVD is derived from the eigendecomposition of a positive semidefinite matrix. This is shown by considering the eigendecomposition of the two positive semidefinite matrices obtained from $\mathbf{A}$, namely, $\mathbf{A}\mathbf{A}^{\mathrm{T}}$ and $\mathbf{A}^{\mathrm{T}}\mathbf{A}$. If we express these matrices in terms of the SVD of $\mathbf{A}$, we find

$$\begin{aligned}\mathbf{A}\mathbf{A}^{\mathrm{T}} &= \mathbf{P}\mathbf{\Delta}\mathbf{Q}^{\mathrm{T}}\mathbf{Q}\mathbf{\Delta}\mathbf{P}^{\mathrm{T}} \\ &= \mathbf{P}\mathbf{\Delta}^{2}\mathbf{P}^{\mathrm{T}}=\mathbf{P}\mathbf{\Lambda}\mathbf{P}^{\mathrm{T}}\end{aligned} \tag{125}$$

and

$$\begin{aligned}\mathbf{A}^{\mathrm{T}}\mathbf{A} &= \mathbf{Q}\mathbf{\Delta}\mathbf{P}^{\mathrm{T}}\mathbf{P}\mathbf{\Delta}\mathbf{Q}^{\mathrm{T}} \\ &= \mathbf{Q}\mathbf{\Delta}^{2}\mathbf{Q}^{\mathrm{T}}=\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{\mathrm{T}}.\end{aligned} \tag{126}$$

This equation shows that $\mathbf{\Delta}$ is the square root of $\mathbf{\Lambda}$, that $\mathbf{P}$ is eigenvectors of $\mathbf{A}\mathbf{A}^{\mathrm{T}}$, and that $\mathbf{Q}$ is eigenvectors of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$.

For example, the matrix

$$\mathbf{A}=\begin{bmatrix}1.1547 & -1.1547\\ -1.0774 & 0.0774\\ -0.0774 & 1.0774\end{bmatrix} \tag{127}$$

can be expressed as

$$\begin{aligned}\mathbf{A} &= \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^{\mathrm{T}} \\ &= \begin{bmatrix} 0.8165 & 0 \\ -0.4082 & -0.7071 \\ -0.4082 & -0.7071 \end{bmatrix} \\ &\quad \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0.7071 & 0.7071 \\ -0.7071 & -0.7071 \end{bmatrix} \\ &= \begin{bmatrix} 1.1547 & -1.1547 \\ -0.0774 & 0.0774 \\ -0.0774 & 1.0774 \end{bmatrix}. \end{aligned} \tag{128}$$

We can check that

$$\begin{aligned}\mathbf{A}\mathbf{A}^{\mathrm{T}} &= \begin{bmatrix} 0.8165 & 0 \\ -0.4082 & -0.7071 \\ -0.4082 & -0.7071 \end{bmatrix} \begin{bmatrix} 2^2 & 0 \\ 0 & 1^2 \end{bmatrix} \\ &\quad \begin{bmatrix} 0.8165 & -0.4082 & -0.4082 \\ 0 & -0.7071 & 0.7071 \end{bmatrix} \\ &= \begin{bmatrix} 2.6667 & 1.3333 & 1.3333 \\ 1.3333 & 1.1667 & 0.1667 \\ 1.3333 & 0.1667 & 1.1667 \end{bmatrix} \end{aligned} \tag{129}$$

and that

$$\begin{aligned}\mathbf{A}^{\mathrm{T}}\mathbf{A} &= \begin{bmatrix} 0.7071 & 0.7071 \\ -0.7071 & 0.7071 \end{bmatrix} \begin{bmatrix} 2^2 & 0 \\ 0 & 1^2 \end{bmatrix} \\ &\quad \begin{bmatrix} 0.7071 & 0.7071 \\ 0.7071 & 0.7071 \end{bmatrix} \\ &= \begin{bmatrix} 2.5 & -1.5 \\ -1.5 & 2.5 \end{bmatrix}. \end{aligned} \tag{130}$$

### *Generalized or Pseudoinverse*

The inverse of a matrix is defined only for full rank square matrices. The generalization of the inverse for other matrices is called *generalized inverse*, *pseudoinverse*, or *Moore–Penrose inverse* and is denoted by $\mathbf{X}^{+}$. The pseudoinverse of $\mathbf{A}$ is the unique matrix that satisfies the following four constraints:

$$\begin{aligned} &\mathbf{A}\mathbf{A}^{+}\mathbf{A} = \mathbf{A} && (i) \\ &\mathbf{A}^{+}\mathbf{A}\mathbf{A}^{+} = \mathbf{A}^{+} && (ii) \\ &(\mathbf{A}\mathbf{A}^{+})^{\mathrm{T}} = \mathbf{A}\mathbf{A}^{+} \quad \text{(symmetry 1)} && (iii) \\ &(\mathbf{A}^{+}\mathbf{A})^{\mathrm{T}} = \mathbf{A}^{+}\mathbf{A} \quad \text{(symmetry 2)} && (iv). \end{aligned} \tag{131}$$

For example, with

$$\mathbf{A} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \\ 1 & 1 \end{bmatrix} \tag{132}$$

we find that the pseudoinverse is equal to

$$\mathbf{A}^{+} = \begin{bmatrix} .25 & -.25 & .5 \\ -.25 & .25 & .5 \end{bmatrix}. \tag{133}$$

This example shows that the product of a matrix and its pseudoinverse does not always give the identity matrix:

$$\begin{aligned}\mathbf{A}\mathbf{A}^{+} &= \begin{bmatrix} 1 & -1 \\ -1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} .25 & -.25 & .5 \\ -.25 & .25 & .5 \end{bmatrix} \\ &= \begin{bmatrix} 0.3750 & 0.1250 \\ 0.1250 & 0.3750 \end{bmatrix} \end{aligned} \tag{134}$$

### *Pseudoinverse and SVD*

The SVD is the building block for the Moore–Penrose pseudoinverse because any matrix $\mathbf{A}$ with SVD equal to $\mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^{\mathrm{T}}$ has for pseudoinverse

$$\mathbf{A}^{+} = \mathbf{Q}\boldsymbol{\Delta}^{-1}\mathbf{P}^{\mathrm{T}}. \tag{135}$$

For the preceding example, we obtain

$$\begin{aligned}\mathbf{A}^{+} &= \begin{bmatrix} 0.7071 & 0.7071 \\ -0.7071 & 0.7071 \end{bmatrix} \begin{bmatrix} 2^{-1} & 0 \\ 0 & 1^{-1} \end{bmatrix} \\ &\quad \begin{bmatrix} 0.8165 & -0.4082 & -0.4082 \\ 0 & -0.7071 & 0.7071 \end{bmatrix} \\ &= \begin{bmatrix} 0.2887 & -0.6443 & 0.3557 \\ -0.2887 & -0.3557 & 0.6443 \end{bmatrix}. \end{aligned} \tag{136}$$

Pseudoinverse matrices are used to solve multiple regression and analysis of variance problems.

*Hervé Abdi and Lynne J. Williams*

***See also*** Analysis of Covariance (ANCOVA); Analysis of Variance (ANOVA); Canonical Correlation Analysis; Confirmatory Factor Analysis; Correspondence Analysis; Discriminant Analysis; General Linear Model; Latent Variable; Mauchly Test; Multiple Regression; Principal Components Analysis; Sphericity; Structural Equation Modeling

**Further Readings**

Abdi, H. (2007). Eigendecomposition: Eigenvalues and eigenvecteurs. In N. J. Salkind (Ed.), *Encyclopedia of measurement and statistics* (pp. 304–308). Thousand Oaks, CA: Sage.

Abdi, H. (2007). Singular value decomposition (SVD) and generalized singular value decomposition (GSVD). In N. J. Salkind (Ed.), *Encyclopedia of measurement and statistics* (pp. 907–912). Thousand Oaks, CA: Sage.

Basilevsky, A. (1983). *Applied matrix algebra in the statistical sciences*. New York: North-Holland.

Graybill, F. A. (1969). *Introduction to matrices with applications in statistics*. Belmont, CA: Wadsworth.

Healy, M. J. R. (1986). *Matrices for statistics*. Oxford, UK: Oxford University Press.

Searle, S. R. (1982). *Matrices algebra useful for statistics*. New York: Wiley.

# Mauchly Test

The Mauchly test (or Mauchly's test) assesses the validity of the sphericity assumption that underlies repeated measures analysis of variance (ANOVA). Developed in 1940 by John W. Mauchly, an electrical engineer who codeveloped the first general-purpose computer, the Mauchly test is the default test of sphericity in several common statistical software programs. Provided the data are sampled from a multivariate normal population, a significant Mauchly test result indicates that the assumption of sphericity is untenable. This entry first explains the sphericity assumption and then describes the implementation and computation of the Mauchly test. The entry ends with a discussion of the test's limitations and critiques.

## The Sphericity Assumption

The sphericity assumption is the assumption that the difference scores of paired levels of the repeated measures factor have equal population variance. As with the other ANOVA assumptions of normality and homogeneneity of variance, it is important to note that the sphericity assumption refers to population parameters rather than sample statistics. Also worth noting is that the sphericity assumption by definition is always met for designs with only two levels of a repeated measures factor. One need not conduct a Mauchly test on such data, and a test conducted automatically by statistical software will not output $p$ values.

Sphericity is a more general form of *compound symmetry*, the condition of equal population covariance (among paired levels) and equal population variance (among levels). Whereas compound symmetry is a sufficient but not necessary precondition for conducting valid repeated measures $F$ tests (assuming normality), sphericity is both a sufficient and necessary precondition. Historically, statisticians and social scientists often failed to recognize these distinctions between compound symmetry and sphericity, leading to frequent confusion over the definitions of both, as well as the true statistical assumptions that underlie repeated measures ANOVA. In fact, Mauchly's definition of sphericity is what is now considered compound symmetry, although the Mauchly test nevertheless assesses what is now considered sphericity.

## Implementation and Computation

Like any null hypothesis significance test, the Mauchly test assesses the probability of obtaining a value for the test statistic as extreme as that observed given the null hypothesis. In this instance, the null hypothesis is that of sphericity, and the test statistic is Mauchly's $W$. Mathematically, the null hypothesis of sphericity (and alternative hypothesis of nonsphericity) can be written in terms of difference scores:

$$H_0 : \sigma^2_{y_1-y_2} = \sigma^2_{y_1-y_3} = \sigma^2_{y_2-y_3} \ldots$$
$$H_1 : \sigma^2_{y_1-y_2} \neq \sigma^2_{y_1-y_3} \neq \sigma^2_{y_2-y_3} \ldots$$

(for all $k(k - 1)/2$ unique difference scores created from $k$ levels of repeated variable $y$) or in terms of matrix algebra:

$$H_0 : \mathbf{C}'\boldsymbol{\Sigma}\,\mathbf{C} = \lambda\mathbf{I}$$
$$H_1 : \mathbf{C}'\boldsymbol{\Sigma}\,\mathbf{C} = \lambda\mathbf{I},$$

where $\mathbf{C}$ is any $(k - 1) \times (k$ orthonormal coefficient matrix associated with the hypothesized repeated measure effect; $\mathbf{C}'$ is the transpose of $\mathbf{C}$; $\boldsymbol{\Sigma}$ is the $k \times k$ population covariance matrix; $\lambda$ is a positive, scalar number; and $\mathbf{I}$ is the $(k - 1) \times (k - 1)$ identity matrix. Mauchly's test statistic, $W$,

can be expressed concisely only in terms of matrix algebra:

$$W = \frac{|\mathbf{C}'\mathbf{S}\mathbf{C}|}{\left[\frac{tr(\mathbf{C}'\mathbf{S}\mathbf{C})}{k-1}\right]^{k-1}},$$

where $\mathbf{S}$ is the $k \times k$ sample covariance matrix.

One can rely on either an approximate or exact sampling distribution to determine the probability value of an obtained $W$ value. Because of the cumbersome computations required to determine exact $p$ values and the precision of the chi-square approximation, even statistical software packages (e.g., SPSS, an IBM company, formerly called PASW© Statistics) typically rely on the latter. The chi-square approximation is based on the statistic $-(n-1)dW$ with degrees of freedom $(df) = k(k-1)/2 - 1$, where

$$d = 1 - \left[\frac{2(k-1)^2 + (k-1) + 2}{6(k-1)(n-1)}\right].$$

For critical values for the exact distribution, see Nagarsenker and Pillai (1973).

## Limitations and Critiques

The Mauchly test is not robust to nonnormality: Small departures from multivariate normality in the population distribution can lead to artificially low or high Type I error (i.e., false positive) rates. In particular, heavy-tailed (leptokurtic) distributions can—under typical sample sizes and significance thresholds—triple or quadruple the number of Type I errors beyond their expected rate. Researchers who conduct a Mauchly test should therefore examine their data for evidence of nonnormality and, if necessary, consider applying normalizing transformations before reconducting the Mauchly test.

Compared with other tests of sphericity, the Mauchly test is not the most statistically powerful. In particular, the *local invariant test* (see Cornell, Young, Seaman, & Kirk, 1992) produces fewer Type II errors (i.e., false negative) than the Mauchly test does. This power difference between the two tests is trivially small for large samples and small $k/n$ ratios but noteworthy for small samples sizes and large $k/n$ ratios. For this reason, some advocate the use of the local invariant test over the Mauchly test.

Finally, some statisticians have called into question the utility of conducting any preliminary test of sphericity such as the Mauchly test. For repeated measures data sets in the social sciences, they argue, sphericity is almost always violated to some degree, and thus researchers should universally correct for this violation (by adjusting *df* with the Greenhouse–Geisser and the Huynh–Feldt estimates). Furthermore, like any significance test, the Mauchly test is limited in its utility by sample size: For large samples, small violations of sphericity often produce significant Mauchly test results, and for small samples, the Mauchly test often does not have the power to detect large violations of sphericity. Finally, critics of sphericity testing note that adoption of the *df* correction tests only when the Mauchly test reveals significant nonsphericity—as opposed to always adopting such *df* correction tests—does not produce fewer Type I or II errors under typical testing conditions (as shown by simulation research).

Aside from using alternative tests of sphericity or forgoing such tests in favor of adjusted *df* tests, researchers who collect data on repeated measures should also consider employing statistical models that do not assume sphericity. Of these alternative models, the most common is multivariate ANOVA (MANOVA). Power analyses have shown that the univariate ANOVA approach possesses greater power than the MANOVA approach when sample size is small ($n < k + 10$) or the sphericity violation is not large ($\varepsilon > .7$) but that the opposite is true when sample sizes are large and the sphericity violation is large.

*Samuel T. Moulton*

***See also*** Analysis of Variance (ANOVA); Bartlett's Test; Greenhouse–Geisser Correction; Homogeneity of Variance; Homoscedasticity; Repeated Measures Design; Sphericity

### Further Readings

Cornell, J. E., Young, D. M., Seaman, S. L., & Kirk, R. E. (1992). Power comparisons of eight tests for sphericity in repeated measures designs. *Journal of Educational Statistics, 17*, 233–249.

Huynh, H., & Mandeville, G. K. (1979). Validity conditions in repeated measures designs. *Psychological Bulletin, 86,* 964–973.
Keselman, H. J., Rogan, J. C., Mendoza, J. L., & Breen, L. J. (1980). Testing the validity conditions of repeated measures *F* tests. *Psychological Bulletin, 87,* 479–481.
Mauchly, J. W. (1940). Significance test for sphericity of a normal *n*-variate distribution. *Annals of Mathematical Statistics, 11,* 204–209.
Nagarsenker, B. N., & Pillai, K. C. S. (1973). The distribution of the sphericity test criterion. *Journal of Multivariate Analysis, 3,* 226–235.

# MBESS

MBESS is an R package that was developed primarily to implement important but nonstandard methods for the behavioral, educational, and social sciences. The generality and applicability of many of the functions contained in MBESS have allowed the package to be used in a variety of other disciplines. Both MBESS and R are open source and freely available from The R Project's Comprehensive R Archive Network. The MBESS Web page contains the reference manual, source code files, and binaries files. MBESS (and R) is available for Apple Macintosh, Microsoft Windows, and Unix/Linux operating systems.

The major categories of functions contained in MBESS are (a) estimation of effect sizes (standardized and unstandardized), (b) confidence interval formation based on central and noncentral distributions ($t$, $F$, and $\chi^2$), (c) sample size planning from the accuracy in parameter estimation and power analytic perspectives, and (d) miscellaneous functions that allow the user to easily interact with R for analyzing and graphing data. Most MBESS functions require only summary statistics. MBESS thus allows researchers to compute effect sizes and confidence intervals based on summary statistics, which facilitates using previously reported information (e.g., for calculating effect sizes to be included in meta-analyses) or if one is primarily using a program other than R to analyze data but still would like to use the functionality of MBESS.

MBESS, like R, is based on a programming environment instead of a point-and-click interface for the analysis of data. Because of the necessity to write code in order for R to implement functions (such as the functions contained within the MBESS package), a resulting benefit is "reproducible research," in the sense that a record exists of the exact analyses performed, with all options and subsamples denoted. Having a record of the exact analyses, by way of a script file, that were performed is beneficial so that the data analyst can (a) respond to inquiries regarding the exact analyses, algorithms, and options; (b) modify code for similar analyses on the same or future data; and (c) provide code and data so that others can replicate the published results. Many novel statistical techniques are implemented in R, and in many ways R has become necessary for cutting-edge developments in statistics and measurement. In fact, R has even been referred to as the lingua franca of statistics.

MBESS, developed by Ken Kelley, was first released publicly in May 2006 and has since incorporated functions contributed by others. MBESS will continue to be developed for the foreseeable future and will remain open source and freely available. Although only minimum experience with R is required in order to use many of the functions contained within the MBESS package, in order to use MBESS to its maximum potential, experience with R is desirable.

*Ken Kelley*

***See also*** Confidence Intervals; Effect Size, Measures of; R; Sample Size Planning

**Further Readings**

de Leeuw, J. (2005). On abandoning XLISP-STAT. *Journal of Statistical Software, 13*(7), 1–5.
Kelley, K. (2006–2008). *MBESS* [computer software and manual]. Accessed February 16, 2010, from http://cran.r-project.org/web/packages/MBESS/index.html
Kelley, K. (2007). Confidence intervals for standardized effect sizes: Theory, application, and implementation. *Journal of Statistical Software, 20*(8), 1–24.
Kelley, K. (2007). Methods for the behavioral, educational, and social science: An R package. *Behavior Research Methods, 39*(4), 979–984.
Kelley, K., Lai, K., & Wu, P.-J. (2008). Using R for data analysis: A best practice for research. In J. Osbourne (Ed.), *Best practices in quantitative methods* (pp. 535–572). Thousand Oaks, CA: Sage.
R Development Core Team. (2008). *The R project for statistical computing*. Retrieved February 16, 2010, from http://www.R-project.org/

Venebles, W. N., Smith, D. M., & The R Development Core Team. (2008). *An introduction to R: Notes on R: A programming environment for data analysis and graphics*. Retrieved February 16, 2010, from http://cran.r-project.org/doc/manuals/R-intro.pdf

**Websites**

Comprehensive R Archive Network: http://cran.r-project.org
MBESS: http://cran.r-project.org/web/packages/MBESS/index.html
The R Project: http://www.r-project.org

# McNemar's Test

McNemar's test, also known as a *test of correlated proportions,* is a nonparametric test used with dichotomous nominal or ordinal data to determine whether two sample proportions based on the same individuals are equal. McNemar's test is used in many fields, including the behavioral and biomedical sciences. In short, it is a test of symmetry between two related samples based on the chi-square distribution with 1 degree of freedom (*df*).

McNemar's test is unique in that it is the only test that can be used when one or both conditions being studied are measured using the nominal scale. It is often used in before–after studies, in which the same individuals are measured at two times, a pretest–posttest, for example. McNemar's test is also often used in matched-pairs studies, in which similar people are exposed to two different conditions, such as a case–control study. This entry details the McNemar's test formula, provides an example to illustrate the test, and examines its application in research.

## Formula

McNemar's test, in its original form, was designed only for dichotomous variables (i.e., yes–no, right–wrong, effect–no effect) and therefore gives rise to proportions. McNemar's test is a test of the equality of these proportions to one another given the fact that they are based in part on the same individual and therefore correlated. More specifically, McNemar's test assesses the significance of any observed change while accounting for the dependent nature of the sample. To do so, a fourfold table of frequencies must be set up to represent the first and second sets of responses from the same or matched individuals. This table is also known as a $2 \times 2$ contingency table and is illustrated in Table 1.

In this table, Cells $A$ and $D$ represent the discordant pairs, or individuals whose response changed from the first to the second time. If an individual changes from $+$ to $-$, he or she is included in Cell $A$. Conversely, if the individual changes from $-$ to $+$, he or she is tallied in Cell $D$. Cells $B$ and $C$ represent individuals who did not change responses over time, or pairs that are in agreement. The main purpose of McNemar's test is determine whether the proportion of individuals who changed in one direction ($+$ to $-$) is significantly different from that of individuals who changed in the other direction ($-$ to $+$).

When one is using McNemar's test, it is unnecessary to calculate actual proportions. The difference between the proportions algebraically and conceptually reduces to the difference between the frequencies given in $A$ and $D$. McNemar's test then assumes that $A$ and $D$ belong to a binomial distribution defined by

$$n = A + D,\ p = .05,\ \text{and}\ q = .05.$$

Based on this, the expectation under the null hypothesis would be that $\frac{1}{2}(A+D)$ cases would change in one direction and $\frac{1}{2}(A+D)$ cases would change in the other direction. Therefore, $Ho: A = D$. The $\chi^2$ formula,

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i},$$

where $O_i =$ observed number of cases in the $i$th category and $E_i =$ expected number of cases in the $i$th category under $H_0$, converts into

$$\chi^2 = \frac{(A - (A+D)/2)^2}{\frac{(A+D)}{2}} + \frac{(D - (A+D)/2)^2}{\frac{(A+D)}{2}}$$

and then factors into

$$\chi^2 = \frac{(A-D)^2}{A+D} \text{ with } df = 1.$$

**Table 1** Fourfold Table for Use in Testing Significance of Change

| | After | |
|---|---|---|
| *Before* | − | + |
| + | A | B |
| − | C | D |

This is McNemar's test formula. The sample distribution is distributed approximately as chi-square with 1 *df*.

### *Correction for Continuity*

The approximation of the sample distribution by the chi-square distribution can present problems, especially if the expected frequencies are small. This is because the chi-square is a continuous distribution whereas the sample distribution is discrete. The correction for continuity, developed by Frank Yates, is a method for removing this source of error. It requires the subtraction of 1 from the absolute value of the difference between $A$ and $D$ prior to squaring. The subsequent formula, including the correction for continuity, becomes

$$\chi^2 = \frac{(|A - D| - 1)^2}{A + D} \text{ with } df = 1.$$

### Small Expected Frequencies

When the expected frequency is very small ($\frac{1}{2}(A + D) < 5$), the binomial test should be used instead.

## Example

Suppose a researcher was interested in the effect of negative political campaign messages on voting behavior. To investigate, the researcher uses a before–after design in which 65 subjects are polled twice on whether they would vote for a certain politician: before and after viewing a negative campaign ad discrediting that politician. The researcher hypothesizes that the negative campaign message will reduce the number of individuals who will vote for the candidate targeted by the negative ad. The data are recorded in the form shown in Table 2. The hypothesis test follows; the data are entirely artificial.

**Table 2** Form of Table to Show Subjects' Change in Voting Decision in Response to Negative Campaign Ad

| | *After Campaign Ad* | |
|---|---|---|
| *Before Campaign Ad* | No Vote | Yes Vote |
| Yes vote | yn | yy |
| No vote | nn | ny |

### *Statistical Test*

McNemar's test is chosen to determine whether there was a significant change in voter behavior. McNemar's test is appropriate because the study uses two related samples, the data are measured on a nominal scale, and the researcher is using a before–after design. McNemar's test formula as it applies to Table 2 is shown below.

$$\chi^2 = \frac{(yn - ny)^2}{yn + ny} \text{ with } df = 1.$$

With the correction for continuity included, the formula becomes

$$\chi^2 = \frac{(|yn - ny| - 1)^2}{yn + ny}.$$

### *Hypotheses*

$H_0$: For those subjects who change, the probability that any individual will change his or her vote from yes to no after being shown the campaign ad (that is, $P_{yn}$) is equal to the probability that the individual will change his or her vote from no to yes (that is, $P_{ny}$), which is equal to $\frac{1}{2}$. More specifically,

$$H_0 : P_{yn} = P_{ny} = \frac{1}{2}.$$

$H_1$: For those subjects who change, the probability that any individual will change his or her vote from yes to no after being shown the negative campaign ad will be significantly greater than the probability that the individual will change his or her vote from no to yes. In other words,

$$H_1 : P_{yn} > P_{ny}.$$

### *Significance Level*

Let $\alpha = .01, N = 65$, the number of individuals polled before and after the campaign ad was shown.

### *Sampling Distribution*

The sampling distribution of $\chi^2$ as computed by McNemar's test is very closely approximated by the chi-square distribution with $df = 1$. In this example, $H_1$ predicts the direction of the difference and therefore requires a one-tailed rejection region. This region consists of all the $\chi^2$ values that are so large they only have a 1% likelihood of occurring if the null hypothesis is true. For a one-tailed test, the critical value with $p < .01$ is 7.87944.

### *Calculation and Decision*

The artificial results of the study are shown in Table 3. The table shows that 30 subjects changed their vote from yes to no ($yn$) after seeing the negative campaign ad, and 7 subjects changed their vote from no to yes ($ny$).

The other two cells, $yy = 11$ and $nn = 17$, represent those individuals who did not change their vote after seeing the ad.

For these data,

$$\chi^2 = \frac{(yn - ny)^2}{yn + ny} = \frac{(30 - 7)^2}{(30 + 7)} = \frac{(23)^2}{37} = 14.30.$$

Including the correction for continuity:

$$\chi^2 = \frac{(|yn - ny - 1)^2|}{yn + ny} = \frac{(|30 - 7| - 1)^2}{30 + 7}$$
$$= \frac{(22)^2}{37} = 13.08.$$

The critical $\chi^2$ value for a one-tailed test at $\alpha = .01$ is 7.87944. Both 14.30 and 13.08 are greater than 7.87944; therefore, the null hypothesis is rejected. These results support the researcher's hypothesis that negative campaign ads significantly decrease the number of individuals willing to vote for the targeted candidate.

**Table 3** Subjects' Voting Decision Before and After Seeing Negative Campaign Ad

| | *After Campaign Ad* | |
|---|---|---|
| *Before Campaign Ad* | No Vote | Yes Vote |
| Yes vote | 30 | 11 |
| No vote | 17 | 7 |

## Application

McNemar's test is valuable to the behavioral and biomedical sciences because it gives researchers a way to test for significant effects in dependent samples using nominal measurement. It does so by reducing the difference between proportions to the difference between discordant pairs and then applying the binomial distribution. It has proven useful in the study of everything from epidemiology to voting behavior, and it has been modified to fit more specific situations, such as misclassified data, improved sample size estimations, multivariate samples, and clustered matched-pair data.

*M. Ashley Morrison*

***See also*** Chi-Square Test; Dichotomous Variable; Distribution; Nominal Scale; Nonparametric Statistics; One-Tailed Test; Ordinal Scale

### Further Readings

Bowker, A. H. (1948). A test for symmetry in contingency tables. *Journal of the American Statistical Association, 43,* 572–574.

Eliasziw, M., & Donner, A. (1991). Application of the McNemar test to non-independent matched pair data. *Statistics in Medicine, 10,* 1981–1991.

Hays, W. L. (1994). *Statistics* (5th ed.). Orlando, FL: Harcourt Brace.

Klingenberg, B., & Agresti, A. (2006). Multivariate extensions of McNemar's test. *Biometrics, 62,* 921–928.

Levin, J. R., & Serlin, R. C. (2000). Changing students' perspectives of McNemar's test of change. *Journal of Statistics Education* [online], *8*(2). Retrieved February 16, 2010, from www.amstat.org/publications/jse/secure/v8n2/levin.cfm

Lyles, R. H., Williamson, J. M., Lin, H. M., & Heilig, C. M. (2005). Extending McNemar's test: Estimation and inference when paired binary outcome data are misclassified. *Biometrics, 61,* 287–294.

McNemar, Q. (1947). Note on sampling error of the difference between correlated proportions or percentages. *Psychometrika, 12,* 153–157.

Satten, G. A., & Kupper, L. L. (1990). Sample size determination for matched-pair case-control studies where the goal is interval estimation of the odds ratio. *Journal of Clinical Epidemiology, 43,* 55–59.

Siegel, S. (1956). *Nonparametric statistics*. New York: McGraw-Hill.
Yates, F. (1934). Contingency tables involving small numbers and the $\chi^2$ test. Supplement to *Journal of the Royal Statistical Society, 1,* 217–235.

# Mean

The mean is a parameter that measures the central location of the distribution of a random variable and is an important statistic that is widely reported in scientific literature. Although the arithmetic mean is the most commonly used statistic in describing the central location of the sample data, other variations of it, such as the truncated mean, the interquartile mean, and the geometric mean, may be better suited in a given circumstance. The characteristics of the data dictate which one of them should be used. Regardless of which mean is used, the sample mean remains a random variable. It varies with each sample that is taken from the same population. This entry discusses the use of mean in probability and statistics, differentiates between the arithmetic mean and its variations, and examines how to determine its appropriateness to the data.

## Use in Probability and Statistics

In probability, the mean is a parameter that measures the central location of the distribution of a random variable. For a real-valued random variable, the mean, or more appropriately the population mean, is the expected value of the random variable. That is to say, if one observes the random variable numerous times, the observed values of the random variable would converge in probability to the mean. For a discrete random variable with a probability function $p(y)$, the expected value exists if

$$\sum_y yp(y) < \infty, \tag{1}$$

where $y$ is the values assigned by the random variable. For a continuous random variable with a probability density function $f(y)$, the expected value exists if

$$\int_{-\infty}^{\infty} |y| f(y) dy < \infty. \tag{2}$$

Comparing Equation 1 with Equation 2, one notices immediately that the $f(y)dy$ in Equation 2 mirrors the $p(y)$ in Equation 1, and the integration in Equation 2 is analogous to the summation in Equation 1.

The above definitions help to understand conceptually the expected value, or the population mean. However, they are seldom used in research to derive the population mean. This is because in most circumstances, either the size of the population (discrete random variables) or the true probability density function (continuous random variables) is unknown, or the size of the population is so large that it becomes impractical to observe the entire population. The population mean is thus an unknown quantity.

In statistics, a sample is often taken to estimate the population mean. Results derived from data are thus called *statistics* (in contrast to what are called *parameters* in populations). If the distribution of a random variable is known, a probability model may be fitted to the sample data. The population mean is then estimated from the model parameters. For instance, if a sample can be fitted with a normal probability distribution model with parameters $\mu$ and $\sigma$, the population mean is simply estimated by the parameter $\mu$ (and $\sigma^2$ as the variance). If the sample can be fitted with a *Gamma distribution* with parameters $\alpha$ and $\beta$, the population mean is estimated by the product of $\alpha$ and $\beta$ (i.e., $\alpha\beta$), with $\alpha\beta^2$ as the variance. For an exponential random variable with parameter $\beta$, the population mean is simply the $\beta$, with $\beta^2$ as the variance. For a chi-square ($\chi^2$) random variable with $\nu$ degrees of freedom, the population mean is $\nu$, with $2\nu$ being the variance.

## Arithmetic Mean

When the sample data are not fitted with a known probability model, the population mean is often inferred from the sample mean, a common practice in applied research. The most widely used sample mean for estimating the population mean is the arithmetic mean, which is calculated as the sum of the observed values of a random variable

divided by the number of observations in the sample.

Formally, for a sample of $n$ observations, $x_1, x_2, \ldots, x_n$ on a random variable $X$, the arithmetic mean ($\overline{x}$) of the sample is defined as

$$\overline{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad (3)$$

where the notation $\sum_{i=1}^{n}$ is a succinct representation of the summation of all values from the first to the last observation of the sample. For example, a sample consisting of five observations with values of 4, 5, 2, 6, and 3 has a mean $4[= (4 + 5 + 2 + 6 + 3)/5]$ according to the above definition. A key property of the mean as defined above is that the sum of deviations from it is zero.

If data are grouped, the sample mean can no longer be constructed from each individual measurement. Instead, it is defined using the midvalue of each group interval ($x_j$) and the corresponding frequency of the group ($f_j$):

$$\overline{x} = \frac{1}{n}\sum_{j=1}^{m} f_j x_j, \qquad (4)$$

where $m$ is the number of groups, and $n$ is the total number of observations in the sample. In Equation 4, $f_j x_j$ is the total value for the $j$th group. A summation of the values of all groups is then the grand total of the sample, which is equivalent to the value obtained through summation, as defined in Equation 3. For instance, a sample of $(n =)$ 20 observations is divided into three groups. The intervals for the three groups are 5 to 9 ($x_1 = 7$), 10 to 14 ($x_2 = 12$), and 15 to 19 ($x_3 = 17$), respectively. The corresponding frequency for each group is $(f_1 =)$ 6, $(f_2 =)$ 5, and $(f_3 =)$ 9. The sample mean according to Equation 4 is then

$$\overline{x} = \frac{7 * 6 + 12 * 5 + 17 * 9}{20} = 12.75.$$

Notice that in Equation 3, we summed up the values of all individual observations before arriving at the sample mean. The summation process is an arithmetic operation on the data. This requires that the data be continuous, that is, they must be either in interval or in ratio scale. For ordinal data, the arithmetic mean is not always the most appropriate measure of the central location; the median is, because it does not require the summation operation.

Notice further that in Equation 3, each observation is given an equal weight. Consequently, the arithmetic mean is highly susceptible to extreme values. Extreme low values would underestimate the mean, while extreme high values would inflate the mean. One must keep this property of the sample arithmetic mean in mind when using it to describe research results.

Because the arithmetic mean is susceptible to variability in the sample data, it is often insufficient to report only the sample mean without also showing the sample standard deviation. Whereas the mean describes the central location of the data, the standard deviation provides information about the variability of the data. Two sets of data with the same sample mean, but drastically different standard deviations, inform the reader that either they come from two different populations or they suffer from variability in quality control in the data collection process. Therefore, by reporting both statistics, one informs the reader of not only the quality of the data but also the appropriateness of using these statistics to describe the data, as well as the appropriate choice of statistical methods to analyze these data subsequently.

## Appropriateness

Whether the mean is an appropriate or inappropriate statistic to describe the data is best illustrated by examples of some highly skewed sample data, such as data on the salaries of a corporation, on the house prices in a region, on the total family income in a nation, and so forth. These types of social economic data are often distorted by a few high-income earners or a few high-end properties. The mean is thus an inappropriate statistic to describe the central location of the data, and the median would be a better statistic for the purpose. On the other hand, if one is interested in describing the height or the test score of students in a school, the sample mean would be a good description of the central tendency of the population as these types of data often follow a unimodal symmetric distribution.

## Variations

Extreme values in a data set, if not inherent in a population, are often erroneous and may have either human or instrumental causes. These so-called outliers are therefore artifacts. In order to better estimate the population mean when extreme values occur in a sample, researchers sometimes order the observations in a sample from the smallest to the largest in value and then remove an equal percentage of observations from both the high end and the low end of the data range before applying the arithmetic mean definition to the sample mean. An example is the awarding of a performance score to an athlete in a sport competition. Both the highest and the lowest scores given by the panel of judges are often removed before a final mean score is awarded. This variation of the arithmetic mean is called the *truncated* (or *trimmed*) *mean*.

Suppose that $n$ observations, $x_1, x_2, \ldots, x_n$, are obtained from a study population and $\alpha$ percentage of data points are removed from either end of the data range. The truncated mean ($\overline{x}_T$) for the sample is then

$$\begin{aligned}\overline{x}_T &= \frac{1}{n(1-2\alpha)}(x_{1+\alpha n} + x_{2+\alpha n} + \cdots + x_{n-\alpha n}) \\ &= \frac{1}{n(1-2\alpha)}\sum_{i=1+\alpha n}^{n-\alpha n} x_i.\end{aligned} \tag{5}$$

In reporting the truncated mean, one must give the percentage of the removed data points in relation to the total number of observations, that is, the value of $\alpha$, in order to inform the reader how the truncated mean is arrived at. Even with the removal of some extreme values, the truncated mean is still not immune to problematic data, particularly if the sample size $n$ is small.

If the entire first quartile and the entire last quartile of the data points are removed after the observations of the data set are ordered from the smallest to the largest in value, the truncated mean of the sample is called the *interquartile mean*. The interquartile mean can be calculated as follows:

$$\overline{x} = \frac{2}{n}\sum_{i=(n/4)+1}^{3n/4} x_i, \tag{6}$$

where the $i$ value beneath $\Sigma$ indicates that the summation starts from the $(n/4+1)$th observation of the data set, and the value above $\Sigma$ signals that the summation ends at the $(3n/4)$th observation. The 2 above $n$ normalizes the interquartile mean to the full $n$ observations of the sample.

The mean is frequently referred to as the *average*. This interchangeable usage sometimes confuses the reader because the median is sometimes also called the *average,* such as what is routinely used in reporting house prices. The reader must be careful about which one of these statistics is actually being referred to.

The arithmetic mean as defined in Equation 3 is not always a good measure of the central location of the sample in some applications. An example is when the data bear considerable variability that has nothing to do with quality control in data collection. Instead, it is inherent in the random process that gives rise to the data, such as the concentration of environmental chemicals in the air. Within a given day at a given location, their concentration could vary in magnitude by multiples. Another example is the growth of bacteria on artificial media. The number of bacteria growing on the media at a given time may be influenced by the number of bacteria on the media at an earlier time, by the amount of media available for growth, by the media type, by the different antibiotics incorporated in the media, by the micro growing environment, and so on. The growth of the bacteria proceeds, not in a linear pattern, but in a multiplicative way. The central tendency of these types of data is best described according to their product, but not their sum. The *geometric mean,* but not the arithmetic mean, would thus be closer to the center of the data values. The geometric mean ($\overline{x}_G$) of a sample of $n$ observations, $x_1, x_2, \ldots, x_n$, is defined as the $n$th root of the product of the $n$ values:

$$\overline{x}_G = \sqrt[n]{x_1 x_2 \ldots x_n} = \left(\prod_{i=1}^{n} x_i\right)^{1/n}. \tag{7}$$

Analogous to the summation representation in Equation 3, the notation $\prod_{i=1}^{n}$ is a succinct

representation of the multiplication of all values of the sample set from the first to the *n*th observation. Therefore, all values from the first to the last observation are included in the product.

Comparing Equation 7 with Equation 3, one can see that the difference is that the geometric mean is obtained by multiplying the observations in the sample first and then taking the *n*th root of their product. In contrast, the arithmetic mean is calculated by adding up the observations first and then dividing their sum by the number of observations in the sample. Because of the multiplying and the taking-the-*n*th-root operations, the geometric mean can be applied only to data of positive values, not to data of negative or zero values.

When the sample size $n$ is large, the product of the values of the observations could be very large, and taking the *n*th root of the product could be difficult, even with modern computers. One way to resolve these difficulties is to transform the value of all observations into a logarithm scale. The multiplication process then becomes a summation process, and the operation of taking the *n*th root of the product is replaced by the division of $n$ from the logarithm sum. The geometric mean is then obtained by applying an antilogarithm operation to the result.

For example, suppose that $n$ observations, $x_1$, $x_2$, . . . , $x_n$, are taken from a random variable. The mean of the logarithmic product of the $n$ values ($\overline{x}_{\log}$) in the sample is

$$\overline{x}_{\log} = \frac{1}{n}\log(x_1 x_2 \ldots x_n) = \frac{1}{n}\sum_{i=1}^{n}\log(x_i), \qquad (8)$$

and the geometric mean is then

$$\overline{x}_G = anti\log(\overline{x}_{\log}). \qquad (9)$$

Here, the base of the logarithm scale can be either $e(= 2.718281828$, the base for natural logarithm) or 10. Most often, 10 is used as the base.

*Shihe Fan*

***See also*** Median; Random Variable; Standard Deviation

**Further Readings**

Bickel, P. J. (1965). On some robust estimates of location. *Annals of Mathematical Statistics, 36,* 847–858.

Foster, D. M. E., & Phillips, G. M. (1984). The arithmetic-harmonic mean. *Mathematics of Computation, 42,* 183–191.

Huber, P. J. (1972). Robust statistics: A review. *Annals of Mathematical Statistics, 43,* 1041–1067.

Ott, L., & Mendenhall, W. (1994). *Understanding statistics* (6th ed.). Pacific Grove, CA: Duxbury.

Prescott, P., & Hogg, R. V. (1977). Trimmed and outer means and their variances. *American Statistician, 31,* 156–157.

Tung, S. H. (1975). On lower and upper bounds of the difference between the arithmetic and the geometric mean. *Mathematics of Computation, 29,* 834–836.

Weerahandi, S., & Zidek, J. V. (1979). A characterization of the general mean. *Canadian Journal of Statistics, 83,* 83–90.

## Mean Comparisons

The term *mean comparisons* refers to the comparison of the average of one or more continuous variables over one or more categorical variables. It is a general term that can refer to a large number of different research questions and study designs. For example, one can compare the mean from one sample of data to a hypothetical population value, compare the means on a single variable from multiple independent groups, or compare the means for a single variable for one sample over multiple measurement occasions. In addition, more complex research designs can employ multiple continuous dependent variables simultaneously, as well as a combination of multiple groups and multiple measurement occasions. Overall, mean comparisons are of central interest in any experimental design and many correlational designs when there are existing categorical variables (e.g., gender).

Two primary questions must be asked in any mean comparison: Are the means statistically different, and how big are the differences? The former question can be answered with a statistical test of the difference in means. The latter is answered with a standardized measure of effect size. Together, these more accurately characterize the nature of mean differences.

## Statistical Differences

Testing for statistical differences attempts to answer the question of whether the observed differences, however large or small, are due to some real effect or simply random sampling error. Depending on the nature of the data and the specific research question at hand, different statistical tests must be employed to properly answer the question of whether there are statistical differences in the means.

### z *Test*

The $z$ test is employed when a researcher wants to answer the question, Is the mean of this sample statistically different from the mean of the population? Here, the researcher would have mean and standard deviation information for the population of interest on a particular continuous variable and data from a single sample on the same variable. The observed mean and standard deviation from the sample would then be compared with the population mean and standard deviation. For example, suppose an organization, as part of its annual survey process, had collected job satisfaction information from all its employees. The organization then wishes to conduct a follow-up study relating to job satisfaction with some of its employees and wishes to make sure the sample drawn is representative of the company. Here, the sample mean and standard deviation on the job satisfaction variable would be compared with the mean and standard deviation for the company as a whole and tested with a $z$ test. This test, however, has limited applications in most research settings, for the simple fact that information for the population is rarely available. When population information is unavailable, different statistical tests must be used.

### t *Test*

Unlike the $z$ test, the $t$ test is a widely used and applicable statistical test. In general, the $t$ test is used to compare two groups on a single continuous dependent variable. Generally, this statistical test answers the question, Is the mean of this sample statistically different from the mean of this other sample? By removing the assumption that there is population information available, the $t$ test becomes a much more flexible statistical technique. The $t$ test can be used with experimental studies to compare two experimental conditions, with correlational studies to compare existing dichotomous groups (e.g., gender), and with longitudinal studies to compare the same sample over two measurement occasions. An important limiting factor of the $t$ test is that it can compare only two groups at a time; investigations of mean differences in more complex research designs require a more flexible analytic technique.

### *Analysis of Variance*

Analysis of variance (ANOVA) is even more flexible than the $t$ test in that it can compare multiple groups simultaneously. Because it is derived from the same statistical model as the $t$ test (i.e., the *general linear model*), ANOVA answers questions similar to those answered by the $t$ test. In fact, when comparisons are between two groups, the $t$ test and ANOVA will yield the exact same conclusions, with the test statistics related by the following formula: $F = t^2$. A significant result from an ANOVA will answer whether at least one group (or condition) is statistically different from at least one other group (or condition) on the continuous dependent variable. If a significant result is found when there are three or more groups or conditions, post hoc tests must be conducted to determine exactly where the significant differences lie.

ANOVA is also the appropriate statistical test to use to compare means when there are multiple categorical independent variables that need to be tested simultaneously. These tests are typically called $n$-way ANOVAs, where the $n$ is the number of independent variables. Here, means among all the conditions are compared. This type of analysis can answer questions of whether there are significant effects for any of the independent variables individually, as well as whether there are any combined interactive effects with two or more of the independent variables. Again, significant results do not reveal the nature of the relationships; post hoc tests must be conducted to determine how the variables relate.

### *Multivariate ANOVA*

Multivariate ANOVA (MANOVA) is a multivariate extension of ANOVA. Like ANOVA, MANOVA can compare the means of two or more groups simultaneously; also, an *n*-way MANOVA, like its *n*-way ANOVA counterpart, can compare the means for multiple independent variables at the same time. The key difference between these two types of analyses is that MANOVA can compare means on multiple continuous dependent variables at the same time, whereas ANOVA can compare means only for a single continuous dependent variable. As such, MANOVA answers the question of whether there are differences between any two groups on any of the dependent variables examined. Effectively, this analysis examines only the question of whether something, on any of the dependent variables, is different between any of the groups examined. As with most of the other analyses, a significant result here will require extensive post hoc testing to determine the precise nature of the differences.

### *Regression*

Mean differences can also be assessed via multiple regression, although this approach is less common. Regression techniques can assess mean differences on a single continuous dependent variable between any number of independent variables with any number of categories per independent variable. When there is a single independent variable with two categories, the statistical conclusions from entering this variable in a regression will exactly equal those from a *t* test; interpreting the sign of the regression weight will determine the nature of the mean differences. When there is a single independent variable with three or more categories, this variable must first be transformed into K − 1 "dummy" variables, where K is the number of categories. Each of these dummy variables is then entered into the regression simultaneously, and the overall statistical conclusions from this model will exactly equal those of ANOVA; however, interpreting magnitude and significance of the regression weights from the regression model will describe the nature of the mean differences, rendering post hoc tests unnecessary. Similarly, creating dummy variables with appropriate multiplicative terms can exactly model the results from an *n*-way ANOVA. Two key points regarding the multiple regression approach are that the statistical results from the regression analysis will always equal the results from the *t* test or ANOVA (because all are derived from the general linear model) and that post hoc tests are generally unnecessary because regression weights provide a way to interpret the nature of the mean differences.

## Effect Sizes

Documenting the magnitude of mean differences is of even greater importance than testing whether two means differ significantly. Psychological theory is advanced further by examining how big mean differences are than by simply noting that two groups (or more) are different. Additionally, practitioners need to know the magnitude of group differences on variables of interest in order to make informed decisions about the use of those variables. Also, psychological journals are increasingly requiring the reporting of effect sizes as a condition of acceptance for publication. As such, appropriate effect sizes to quantify mean differences need to be examined. Generally, two types of effect size measures are commonly used to compare means: mean differences and correlational measures.

### *Mean Differences*

Mean difference effect size measures are designed specifically to compare two means at a time. As such, they are very amenable to providing an effect size measure when one is comparing two groups via a *z* test or a *t* test. Mean difference effect sizes can also be used when one is comparing means with an ANOVA, but each condition needs to be compared with a control group or focal group. Because the coding for a categorical variable is arbitrary, the sign of the mean difference effect size measure is also arbitrary; in order for effect sizes to be interpretable, the way in which groups are specified must be very clear. Three common effect size measures are the simple mean difference, the standardized mean difference, and the standardized mean difference designed for experimental studies.

#### Simple Mean Difference

The simplest way to quantify the magnitude of differences between groups on a single continuous dependent variable is to compute a simple mean difference, $M_1 - M_2$, where $M_1$ and $M_2$ are the means for Groups 1 and 2, respectively. Despite its simplicity, this measure has several shortcomings. The most important limitation is that the simple mean difference is scale and metric dependent, meaning that simple mean differences cannot be directly compared on different scales, or even on the same scale if the scale has been scored differently. As such, simple mean differences are best used when there is a well-established scale whose metric does not change. For example, simple mean differences can be used to compare groups on the SAT (a standardized test for college admissions) because it is a well-established test whose scoring does not change from year to year; however, simple mean differences on the SAT cannot be compared with simple differences on the ACT, because these two tests are not scored on the same scale.

#### Standardized Mean Difference

The standardized mean difference, also known as Cohen's *d,* addresses the problem of scale or metric dependence by first standardizing the means to a common metric (i.e., a standard deviation metric, where the *SD* equals 1.0). The equation to compute the standardized mean difference is

$$d = \frac{M_1 - M_2}{\sqrt{\frac{N_1(SD_1)^2 + N_2(SD_2)^2}{N_1 + N_2}}}, \qquad (1)$$

where $N$ is the sample size, $SD$ is the standard deviation, and other terms are defined as before. Examination of this equation reveals that the numerator is exactly equal to the simple mean difference; it is the denominator that standardizes the mean difference by dividing by the pooled within-group standard deviation. Generally, this measure of effect size is preferred because it can be compared directly across studies and aggregated in a meta-analysis. Also, unlike some other measures of mean differences, it is generally insensitive to differences in sample sizes between groups. There are several slight variants to Equation 1 to address some statistical issues in various experimental and correlational settings, but each is very closely related to Equation 1.

Another advantage of the standardized mean difference is that, because it is on a standardized metric, some rough guidelines for interpretation of the numbers can be provided. As a general rule, standardized mean differences (i.e., $d$ values) of $d = 0.20$ are considered small, $d = 0.50$ are medium, and $d = 0.80$ are considered to be large. Of course, the meaning of the mean differences must be interpreted from within the research context, but these guidelines provide a rough metric with which to evaluate the magnitude of mean differences obtained via Equation 1.

#### Standardized Mean Difference for Experiments

In experimental studies, it is often the case that the experimental manipulation will shift the mean of the experimental group high enough to run into the top of the scale (i.e., create a ceiling effect); this event decreases the variability of the experimental group. As such, using a pooled variance will actually underestimate the expected variance and overestimate the expected effect size. In this case, researchers often use an alternative measure of the standardized mean difference,

$$d = \frac{M_{\text{Exp}} - M_{\text{Con}}}{SD_{\text{Con}}}, \qquad (2)$$

where the Con and Exp subscripts denote the control and experimental subgroups, respectively. This measure of effect size can be interpreted with the same metric and same general guidelines as the standardized mean difference from Equation 1.

### *Correlational Effect Sizes*

Correlational-type effect sizes are themselves reported on two separate metrics. The first is on the same metric as the correlation, and the second is on the variance-accounted-for metric (i.e., the correlation-squared metric, or $R^2$). They can be converted from one to the other by a simple square or square-root transformation. These types of effect sizes answer questions about the magnitude

of the relationship between group membership and the dependent variable (in the case of correlational effect sizes) or the percentage of variance accounted for in the dependent variable by group membership. Most of these effect sizes are insensitive to the coding of the group membership variable. The three main types of effect sizes are the point-biserial correlation, the eta (or eta-squared) coefficient, and multiple $R$ (or $R^2$).

#### Point-Biserial Correlation

Of the effect sizes mentioned here, the point-biserial correlation is the only correlational effect size whose sign is dependent on the coding of the categorical variable. It is also the only measure presented here that requires that there be only two groups in the categorical variable. The equation to compute the point-biserial correlation is

$$r_{pb} = \frac{M_1 - M_2}{SD_{\text{Tot}}} \sqrt{\frac{n_1 n_2}{N(N-1)}}, \qquad (3)$$

where $SD_{\text{Tot}}$ is the total standard deviation across all groups; $n_1$ and $n_2$ are the sample sizes for Groups 1 and 2, respectively; and $N$ is total sample size across the two groups. Though it is a standard Pearson correlation, it does not range from zero to $\pm 1.00$; the maximum absolute value is about 0.78. The point-biserial correlation is also sensitive to the proportion of people in each group; if the proportion of people in each group differs substantially from 50%, the maximum value drops even further away from 1.00.

#### Eta

Although the eta coefficient can be interpreted as a correlation, it is not a form of a Pearson correlation. While the correlation is a measure of the linear relationship between variables, the eta actually measures any relationship between the categorical independent variable and the continuous dependent variable. Eta-squared is the square of the eta coefficient and is the ratio of the between-group variance to the total variance. The eta (and eta-squared) can be computed with any number of independent variables and any number of categories in each of those categorical variables. The eta tells the magnitude of the relationship between the categorical variable(s) and the dependent variable but does not describe it; post hoc examinations must be undertaken to understand the nature of the relationship. The eta-squared coefficient tells the proportion of variance that can be accounted for by group membership.

#### Multiple R

The multiple $R$ or $R^2$ is the effect size derived from multiple regression techniques. Like a correlation or the eta coefficient, the multiple $R$ tells the magnitude of the relationship between the set of categorical independent variables and the continuous dependent variable. Also similarly, the $R^2$ is the proportion of variance in the dependent variable accounted for by the set of categorical independent variables. The magnitude for the multiple $R$ (and $R^2$) will be equal to the eta (and eta$^2$) for the full ANOVA model; however, substantial post hoc tests are unnecessary in a multiple regression framework, because careful interpretation of the regression weights can describe the nature of the mean differences.

## Additional Issues

As research in the social sciences increases at an exponential rate, cumulating research findings across studies becomes increasingly important. In this context, knowing whether means are statistically different becomes less important, and documenting the magnitude of the difference between means becomes more important. As such, the reporting of effect sizes is imperative to allow proper accumulation across studies. Unfortunately, current data accumulation (i.e., meta-analytic) methods require that a single continuous dependent variable be compared on a single dichotomous independent variable. Fortunately, although multiple estimates of these effect sizes exist, they can readily be converted to one another. In addition, many of the statistical tests can be converted to an appropriate effect size measure.

Converting between a point-biserial correlation and a standardized mean difference is relatively easy if one of them is already available. For example, the formula for the conversion of a point-biserial correlation to a standardized mean difference is

$$d = \frac{r_{pb}}{\sqrt{p_1 p_2 (1 - r_{pb}^2)}}, \quad (4)$$

where $d$ is the standardized mean difference, $r_{pb}$ is the point-biserial correlation, and $p_1$ and $p_2$ are the proportions in Groups 1 and 2, respectively. The reverse of this formula, for the conversion of a standardized mean difference to a point-biserial correlation, is

$$r_{pb} = \frac{d}{\sqrt{d^2 + \frac{1}{p_1 p_2}}}, \quad (5)$$

where terms are defined as before. If means, standard deviations, and sample sizes are available for each of the groups, then these effect sizes can be computed with Equations 1 through 3. Eta coefficients and multiple $R$s cannot be converted readily to a point-biserial correlation or a standardized mean difference unless there is a single dichotomous variable; then these coefficients equal the point-biserial correlation.

Most statistical tests are not amenable to ready conversion to one of these effect sizes; $n$-way ANOVAs, MANOVAs, regressions with more than one categorical variable, and regressions and ANOVAs with a single categorical variable with three or more categories do not convert to either the point-biserial correlation or the standardized mean difference. However, the $F$ statistic from an ANOVA with two categorical variables is exactly equal to the value of the $t$ statistic via the relationship $F = t^2$. To convert from a $t$ statistic to a point-biserial correlation, the following equation must be used:

$$r_{pb} = \sqrt{\frac{t^2}{t^2 + df}}, \quad (6)$$

where $t$ is the value of the $t$ test and $df$ is the degrees of freedom for the $t$ test. The point-biserial correlation can then be converted to a standardized mean difference if necessary.

*Matthew J. Borneman*

***See also*** Analysis of Variance (ANOVA); Cohen's $d$ Statistic; Effect Size, Measures of; Multiple Regression; Multivariate Analysis of Variance (MANOVA); $t$ Test, Independent Samples;

**Further Readings**

Bonnett, D. G. (2008). Confidence intervals for standardized linear contrasts of means. *Psychological Methods, 13,* 99–109.

Bonnett, D. G. (2009). Estimating standardized linear contrasts of means with desired precision. *Psychological Methods, 14,* 1–5.

Fern, E. F., & Monroe, K. B. (1996). Effect size estimates: Issues and problems in interpretation. *Journal of Consumer Research, 23,* 89–105.

Keselman, H. J., Algina, J., Lix, L. M., Wilcox, R. R., & Deering, K. N. (2008). A generally robust approach for testing hypotheses and setting confidence intervals for effect sizes. *Psychological Methods, 13,* 110–129.

McGrath, R. E., & Meyer, G. J. (2006). When effect sizes disagree: The case of $r$ and $d$. *Psychological Methods, 11,* 386–401.

McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin, 111,* 361–365.

Ruscio, J. (2008). A probability-based measure of effect size: Robustness to base rates and other factors. *Psychological Methods, 13,* 19–30.

# Median

The median is one of the location parameters in probability theory and statistics. (The others are the mean and the mode.) For a real valued random variable $X$ with a cumulative distribution function $F$, the median of $X$ is the unique number that satisfies $F(-m) \leq \frac{1}{2} \geq F(m)$. In other words, the median is the number that separates the upper half from the lower half of a population or a sample. If a random variable is continuous and has a probability density function, half of the area under the probability density function curve would be to the left of $m$ and the other half to the right of $m$. For this reason, the median is also called the 50th percentile (the $i$th percentile is the value such that $i$% of the observations are below it). In a box plot (also called box-and-whisker plot), the median is the central line between the lower and the higher hinge of the box. The location of this central line suggests the central tendency of the underlying data.

The population median, like the population mean, is generally unknown. It must be inferred from the sample median, just like the use of the sample mean for inferring the population mean. In circumstances in which a sample can be fitted with a known probability model, the population median may be obtained directly from the model parameters. For instance, a random variable that follows an exponential distribution with a scale parameter $\beta$ (a scale parameter is the one that stretches or shrinks a distribution), the median is $\beta\ln 2$ (where ln means natural logarithm, which has a base $e = 2.718281828$). If it follows a normal distribution with a location parameter $\mu$ and a scale parameter $\sigma$, the median is $\mu$. For a random variable following a Weibull distribution with a location parameter $\mu$, a scale parameter $\alpha$, and a shape parameter $\gamma$ (a shape parameter is the one that changes the shape of a distribution), the median is $\mu + \alpha(\ln 2)^{1/\gamma}$. However, not all distributions have a median in closed form. Their population median cannot be obtained directly from a probability model but has to be estimated from the sample median.

## Definition and Calculation

The sample median can be defined similarly, irrespective of the underlying probability distribution of a random variable. For a sample of $n$ observations, $x_1, x_2, \ldots x_n$, taken from a random variable $X$, rank these observations in an ascending order from the smallest to the largest in value; the sample median, $m$, is defined as

$$m = \begin{cases} x_k & \text{if } n = 2k+1 \\ (x_k + x_{k+1})/2 & \text{if } n = 2k \end{cases}. \qquad (1)$$

That is, the sample median is the value of the middle observation of the ordered statistics if the number of observations is odd or the average of the value of the two central observations if the number of observations is even. This is the most widely used definition of the sample median.

According to Equation 1, the sample median is obtained from order statistics. No arithmetical summation is involved, in contrast to the operation of obtaining the sample mean. The sample median can therefore be used on data in ordinal, interval, and ratio scale, whereas the sample mean is best used on data in interval and ratio scale because it requires first the summation of all values in a sample.

The sample median as defined in Equation 1 is difficult to use when a population consists of all integers and a sample is taken with an even number of observations. Because the median should also be an integer, two medians could result. One may be called the lower median and the other, the upper median. To avoid calling two medians of a single sample, an alternative is simply to call the upper median the sample median, ignoring the lower one.

If the sample data are grouped into classes, the value of the sample median cannot be obtained according to Equation 1 as the individual values of the sample are no longer available. Under such a circumstance, the median is calculated for the particular class that contains the median. Two different approaches may be taken to achieve the same result. One approach starts with the frequency and cumulative frequency (see Ott and Mendenhall, 1994):

$$m = L + \frac{w}{f_m}\left(\frac{n}{2} - cf_b\right), \qquad (2a)$$

where $m$ = the median, $L$ = lower limit of the class that contains the median, $n$ = total number of observations in the sample, $cf_b$ = cumulative frequency for all classes before the class that contains the median, $f_m$ = frequency for the class that contains the median, and $w$ = interval width of the classes.

The other approach starts with the percentage and cumulative percentage:

$$m = L + \frac{w}{P_m}(50 - cP_b), \qquad (2b)$$

where 50 = the 50th percentile, $cP_b$ = cumulative percentage for all classes before the class that contains the median, and $P_m$ = percentage of the class that contains the median. Both $L$ and $w$ are defined as in Equation 2a. A more detailed description of this approach can be found in *Arguing With Numbers* by Paul Gingrich.

This second approach is essentially a special case of the approach used to interpolate the distance to a given percentile in grouped data. To do so, one needs only to replace the 50 in Equation 2b with a percentile of interest. The percentile

within an interval of interest can then be interpolated from the lower percentile bound of the interval width.

To show the usage of Equations 2a and 2b, consider the scores of 50 participants in a hypothetical contest, which are assigned into five classes with the class interval width = 20 (see Table 1). A glance at the cumulative percentage in the rightmost column of the table suggests that the median falls in the 61-to-80 class because it contains the 50th percentile of the sample population. According to Equation 2a, therefore, $L = 61$, $n = 50$, $cf_b = 13$, $f_m = 20$, and $w = 20$. The interpolated value for the median then is

$$m = 61 + \frac{20 \times (50/2 - 13)}{20} = 73.$$

Using Equation 2b, we have $cP_b = 26$, $P_m = 40$, both $L = 61$, and $w = 20$, as before. The interpolated value for the median of the 50 scores, then, is

$$m = 61 + \frac{20 \times (50 - 26)}{40} = 73.$$

Equations 2a and 2b are equally applicable to ordinal data. For instance, in a survey on the quality of customer services, the answers to the customer satisfaction question may be scored as *dissatisfactory, fairly satisfactory, satisfactory,* and *strongly satisfactory.* Assign a value of 1, 2, 3, or 4 (or any other ordered integers) to represent each of these classes from *dissatisfactory* to *strongly satisfactory* and summarize the number of responses corresponding to each of these classes in a table similar to the one above. One can then apply either Equation 2a or Equation 2b to find the class that contains the median of the responses.

**Table 1** Grouped Data for 50 Hypothetical Test Scores

| | *Frequency* | | | |
|---|---|---|---|---|
| *Class* | Number of Observations | Cumulative Number of Observations | % | Cumulative % |
| 0–20 | 2 | 2 | 4 | 4 |
| 21–40 | 5 | 7 | 10 | 14 |
| 41–60 | 6 | 13 | 12 | 26 |
| 61–80 | 20 | 33 | 40 | 66 |
| 81–100 | 17 | 50 | 34 | 100 |
| | 50 | | 100 | |

## Determining Which Central Tendency Measure to Use

As pointed out at the beginning, the mean, the median, and the mode are all location parameters that measure the central tendency of a sample. Which one of them should be used for reporting a scientific study? The answer to this question depends on the characteristics of the data, or more specifically, on the *skewness* (i.e., the asymmetry of distribution) of the data. A distribution is said to be skewed if one of its two tails is longer than the other. In statistics literature, the mean ($\mu$), median (m), and mode (M) inequality are well known for both continuous and discrete unimodal distributions. The three statistics occur either in the order of $M \leq m \leq \mu$ or in a reverse order of $M \geq m \geq \mu$, depending on whether the random variable is positively or negatively skewed.

For random variables that follow a symmetric distribution such as the normal distribution, the sample mean, median, and mode are equal and can all be used to describe the sample central tendency. Despite this, the median, as well as the mode, of a normal distribution is not used as frequently as the mean. This is because the variability ($V$) associated with the sample mean is much smaller than the $V$ associated with the sample median ($V[m] = [1.2533]^2 V[\mu]$). If a random variable follows a skewed (i.e., nonsymmetrical) distribution, the sample mean, median, and mode are not equal. The median differs substantially from both the arithmetic mean and the mode and is a better measure of the central tendency of a random sample because the median is the minimizer of the mean absolute deviation in a sample. Take a sample set {2, 2, 3, 3, 3, 4, 15} as an example. The median is 3 (so is the mode), which is a far better measure of the centrality of the data set than the arithmetic mean of 4.57. The latter is largely influenced by the last extreme value, 15, and does not adequately describe the central tendency of the data set. From this simple illustration, it can be concluded that the sample median should be favored

over the arithmetic sample mean in describing the centrality whenever the distribution of a random variable is skewed. Examples of such skewed data can be found frequently in economic, sociological, education, and health studies. A few examples are the salary of employees of a large corporation, the net income of households in a city, the house price in a country, and the survival time of cancer patients. A few high-income earners, or a few high-end properties, or a few longer survivors could skew their respective sample disproportionally. Use of the sample mean or mode to represent the data centrality would be inappropriate.

The median is sometimes called the *average*. This term may be confused with the mean for some people who are not familiar with a specific subject in which this interchangeable usage is frequent. In scientific reporting, this interchangeable use is better avoided.

### Advantages

Compared with the sample mean, the sample median has two clear advantages in measuring the central tendency of a sample. The first advantage is that the median can be used for all data measured in ordinal, interval, and ratio scale because it does not involve the mathematic operation of summation, whereas the mean is best used for data measured in interval and ratio scale. The second advantage is that the median gives a measure of central tendency that is more robust than the mean if outlier values are present in the data set because it is not affected by whether the distribution of a random variable is skewed. In fact, the median, not the mean, is a preferred parameter in describing the central tendency of such random variables when their distribution is skewed. Therefore, whether to use the sample median as a central tendency measure depends on the data type. The median is used if a random variable is measured in ordinal scale or if a random variable produces extreme values in a set. In contrast, the mean is a better measure of the sample central tendency if a random variable is continuous and is measured in interval or ratio scale, and if data arising from the random variable contain no extreme value.

When one is using the sample median, it helps to remember its four important characteristics, as pointed out by Lyman Ott and William Mendenhall:

1. The median is the central value of a data set, with half of the set above it and half below it.
2. The median is between the largest and the smallest value of the set.
3. The median is free of the influence of extreme values of the set.
4. Only one median exists for the set (except in the difficult case in which an even number of observations is taken from a population consisting of only integers).

*Shihe Fan*

***See also*** Central Tendency, Measures of; Mean; Mode

**Further Readings**

Abdous, B., & Theodorescu, R. (1998). Mean, median, mode IV. *Statistica Neerlandica, 52,* 356–359.

Gingrich, P. (1995). *Arguing with numbers: Statistics for the social sciences*. Halifax, Nova Scotia, Canada: Fernwood.

Groneveld, A., & Meeden, G. (1977). The mode, median, and mean inequality. *American Statistician, 31,* 120–121.

Joag-Dev, K. (1989). MAD property of a median: A simple proof. *American Statistician, 43,* 26–27.

MacGillivray, H. L. (1981). The mean, median, mode inequality and skewness for a class of densities. *Australian Journal of Statistics, 23,* 247–250.

Ott, L., & Mendenhall, W. (1994). *Understanding statistics* (6th ed.). Pacific Grove, CA: Duxbury.

Wackerly, D. D, Mendenhall, W., III, & Scheaffer, R. L. (2002). *Mathematical statistics with applications* (6th ed.). Pacific Grove, CA: Duxbury.

## Meta-Analysis

Meta-analysis is a statistical method that integrates the results of several independent studies considered to be "combinable." It has become one of the major tools to integrate research findings in social and medical sciences in general and in education and psychology in particular. Although the history of meta-analytic procedures goes all the way back

to the early 1900s and the work of Karl Pearson and others, who devised statistical tools to compare studies from different samples, Gene V. Glass coined the term in 1976. Glass, Barry McGaw, and Mary Lee Smith described the essential characteristics of meta-analysis as follows:

1. It is undeniably quantitative, that is, it uses numbers and statistical methods for organizing and extracting information.
2. It does not prejudge research findings in terms of research quality (i.e., no a priori arbitrary and nonempirical criteria of research quality are imposed to exclude a large number of studies).
3. It seeks general conclusions from many separate investigations that address related or identical hypotheses.

Meta-analysis involves developing concise criteria for inclusion (i.e., *sampling*), searching the literature for relevant studies (i.e., *recruitment*), coding study variables (i.e., *data entry*), calculating standardized effect sizes for individual studies, and generating an overall effect size across studies (i.e., *data analysis*). Unlike primary studies, in which each case in a sample is a unit of analysis, the unit of analysis for meta-analysis is the individual study. The effect sizes calculated from the data in an individual study are analogous to the dependent variable, and the substantive and methodological characteristics affecting the study results are defined as independent variables. Any standardized index that can be used to understand different statistical findings across studies in a common metric can be used as an "effect size." The effect size metric represents both the magnitude and direction of the relation of interest across different primary studies in a standardized metric. A variety of alternatives are available for use with variables that are either continuous or discrete, such as the accumulation of correlations (*effect size r*), and standardized differences between mean scores (effect size d), $p$ values, or $z$ scores effect size (ES). The dependent variable in meta-analysis is computed by transforming findings of each reviewed study into a common metric that relies on either $r$ or $d$ as the combined statistic.

Meta-analysis is not limited to descriptive reviews of research results but can also examine how and why such findings occur. With the use of multivariate statistical applications, meta-analysis can address multiple hypotheses. It may examine the relation between several variables and account for consistencies as well as inconsistencies within a sample of study findings. Because of demand for robust research findings and with the advance of statistical procedures, meta-analysis has become one of the major tools for integrating research findings in social and medical science as well as the field of education, where it originated. A recent search of the ERIC database identified more than 618 articles published between 1980 and 2000 that use *meta-analysis* in their title, as opposed to only 36 written before 1980. In the field of psychology, the gap was 12 versus 1,623, and in the field of medical studies, the difference is even more striking: 7 versus 3,571. Evidence in other fields shows the same trend toward meta-analysis's becoming one of the main tools for evidence-based research.

According to the publication manual of the American Psychological Association, a review article organizes, integrates, and critically evaluates already published material. Meta-analysis is only one way of reviewing or summarizing research literature. Narrative review is the more traditional way of reviewing research literature. There are several differences between traditional narrative reviews and meta-analysis. First, because there are very few systematic procedures, the narrative review is more susceptible to subjective bias and therefore more prone to error than are meta-analytic reviews. In the absence of formal guidelines, reviewers of a certain literature can disagree about many critical issues, such as which studies to include and how to support conclusions with a certain degree of quantitative evidence. In an adequately presented meta-analytical study, one should be able to replicate the review by following the procedure reported in the study.

Narrative review and meta-analysis are also different in terms of the scope of the studies that they can review. The narrative review can be inefficient for reviewing 50 or more studies. This is especially true when the reviewer wants to go beyond describing the findings and explain multiple relations among different variables. Unlike narrative reviews, meta-analysis can put together all available data to answer questions about overall study findings and how they can be accounted for by

various factors, such as sample and study characteristics. Meta-analysis can, therefore, lead to the identification of various theoretical and empirical factors that may permit a more accurate understanding of the issues being reviewed. Thus, although meta-analysis can provide a better assessment of literature because it is more objective, replicable, and systematic, it is important to note that a narrative description for each study is key to any good meta-analytic review as it will help a meta-analyst determine which studies to include and what qualitative information about the studies can and should be coded and statically related to quantitative outcomes in order to evaluate the complexity of topics being reviewed.

The remainder of this entry addresses the methodological issues associated with meta-analytic research and then describes the steps involved in conducting meta-analysis.

## Methodological Issues

Like any other research strategy, meta-analysis is not a perfect solution in research review. Glass summarized the main issues with meta-analysis in four domains: quality, commensurability, selection bias, and nonindependence. The *quality problem* has been a very controversial issue in meta-analysis. At issue is whether the quality of studies be included as a selection criteria. To avoid any bias in selection one option is to include as many studies as possible, regardless of their quality. Others, however, question the practice of including studies of poor quality as doing so limits the validity of the overall conclusions of the review. The *commensurability problem* refers to the most common criticism of meta-analysis: that it compares apples and oranges. In other words, meta-analysis is illogical because it mixes constructs from studies that are not the same.

The *selection-bias problem* refers to the inevitable scrutiny of the claim that the meta-analytic review is comprehensive and nonbiased in its reviewing process. Meta-analysis is not inherently immune from selection bias, as its findings will be biased if there are systematic differences across journal articles, book articles, and unpublished articles. Publication bias is a major threat to the validity of meta-analysis. The *file drawer effect,* which refers to the fact that published studies are less likely to show statistically nonsignificant results, may introduce a bias in the conclusions of meta-analyses. Not only can the decision of whether to include unpublished studies lead to bias, but decisions about how to obtain data and which studies to include can also contribute to selection bias. The unpublished studies that can be located may thus be an unrepresentative sample of unpublished studies. A review of meta-analytical studies published between 1988 and 1991 indicated that most researchers had searched for unpublished material, yet only 31% included unpublished studies in their review. Although most of these researchers supported the idea of including unpublished data in meta-analysis, only 47% of journal editors supported this practice.

The *nonindependence problem* in meta-analysis refers to the assumption that each study in the review is taken randomly from a common population; that the individual studies are independent of one another. "Lumping" sets of independent studies can reduce the reliability of estimations of averages or regression equations. Although Glass and his associates argued that the nonindependence assumption is a matter of practicality, they admit that this problem is the one criticism that is not "off the mark and shallow" (Glass et al., p. 229).

## Steps

Although meta-analytic reviews can take different forms depending on the field of study and the focus of the review, there are five general steps in conducting meta-analysis. The first step involves defining and clarifying the research question, which includes selecting inclusion criteria. Similar to selecting a sample for an empirical study, inclusion criteria for a meta-analysis have to be specified following a theoretical or empirical guideline. The inclusion criteria greatly effect the conclusions drawn from a meta-analytic review. Moreover, the inclusion criteria are one of the steps in a meta-analytic study where bias or subjectivity comes into play. Two critical issues should be addressed at this stage of meta-analysis: (a) Should unpublished studies be included? and (b) should the quality of the studies be included as part of the inclusion criteria? There are no clear answers to these questions. Glass and colleagues, for example, argued against strict inclusion criteria based on

assessing study quality a priori because a meta-analysis itself can empirically determine whether study quality is related to variance in reported study findings. While Glass and others argued for inclusion of all studies, including unpublished reports in order to avoid publication bias toward null findings in the literature, it is possible to empirically assess research quality with a set of methodological variables as part of the meta-analytic data analysis. In other words, instead of eliminating a study based on the reviewer's judgment of its quality, one can empirically test the impact of study quality as a control or moderator variable.

The next step in meta-analysis is to identify studies to be included in the review. This step involves a careful literature search that involves computerized and manual approaches. Computerized search approaches include using discipline specific databases such as PsycINFO in psychology, ERIC in education, MEDLINE in medical sciences, or Sociological Abstracts in sociology. Increasingly, searching the Internet with search engines such as Google (or Google Scholar) also helps identify relevant studies for meta-analytic review. All databases must be searched with the same set of keywords and search criteria in order to ensure reliability across the databases. It is also important to keep in mind that several vendors market the most popular databases, such as PsycINFO, and each vendor has a different set of defaults that determine the outcome of any search. It is, therefore, advisable for investigators to generate a single yet detailed logical search code and test it by using various vendors to see if their databases yield the same result.

Although computerized search engines save time and make it possible to identify relevant materials in large databases, they should be complemented with additional search strategies, including manual search. In fields in which there is no universally agreed-on keyword, for example, one can search key publications or citations of classic articles using the Social Science Citation Index, which keeps track of unique citations of each published article. If narrative reviews have been published recently, one can also check the cited articles in those reviews. Finally, once the final review pool is determined, one must also manually check the references in each of the articles to see if there are relevant studies that have not yet been included in the final pool. Each of these postelectronic search steps can also serve as a reliability check to see whether the original search code works well. In other words, if there are too many articles that were not part of the electronically searched pool, then it is possible that the search code was not a valid tool to identify relevant studies for the review. In those circumstances a modified search would be in order.

The third step in meta-analysis is the development of a coding schema. The goal of study coding is to develop a systematic procedure for recording the appropriate data elements from each study. William A. Stock identified six categories of study elements for systematic coding that address both substantive and methodological characteristics: *report identification* (study identifiers such as year of publication, authors), *setting* (the location or context of the study), *subjects* (participant characteristics), *methodology* (research design characteristics), *treatment* (procedures), and *effect size* (statistical data needed to calculate common effect size). One can modify these basic categories according to the specific focus of the review and with attention to the overall meta-analytic question and potential moderator factors. To further refine the coding scheme, a small subsample of the data ($k = 10$) must be piloted with two raters who did not take part in the creation of the coding schema.

The next step is to calculate effect sizes for each study by transforming individual study statistics into a common effect size metric. The goal of effect size transformation is to reflect with a common metric the relative magnitude of the relations reported in various independent studies. The three most commonly used effect size metrics in meta-analytic reviews are *Cohen's d, correlation coefficient r,* and *odds ratio*. Cohen's *d,* or *effect size d,* is a metric that is used when the research involves mean differences or group contrasts. This is a method used in treatment studies or any design that calls for calculating standardized mean differences across groups in a variable that is continuous in nature. Correlation coefficient *r* can also serve as an effect size metric (or *effect size r*) when the focus of the review is identification of the direction and magnitude of the association between variables. Odds-ratio effect size is commonly used in

epidemiological reviews or in reviews that involve discontinuous variables (e.g., school dropout or diagnosis of a certain condition).

The calculation of an effect size index also requires a decision about the unit of analysis in a meta-analysis. There are two alternatives. The first alternative is to enter the effect size for each variable separately. For example, if a study reports one correlation on the basis of grade point average and another correlation on the basis of an achievement test score, there will be two different effect sizes for the study, one for grade point average and the other for achievement test score. Similarly, if correlations were reported for girls and boys separately, there will be two effect sizes, one for girls and one for boys. The second alternative is to use each study as the unit of analysis. This can be done by averaging effect sizes across the groups. For example, one could take the mean of the correlations for girls and boys and report a single effect size. Both of these approaches have their shortcomings. The former approach gives too much weight to those studies that have more outcome measures, but the latter approach obscures legitimate theoretical and empirical differences across dependent measures (i.e., gender differences may serve as a moderator in certain meta-analytic reviews).

Mark W. Lipsey and David B. Wilson suggest a third alternative that involves calculating an effect size for each independent sample when the focus of analysis is the *sample characteristics* (e.g., age, gender, race) but allowing for multiple effect sizes from a given study when the focus of the analysis is the *study characteristics* (e.g., multiple indicators of the same construct). In other words, the first alternative can be used to calculate an effect size for each distinct construct in a particular study; this alternative yields specific information for each particular construct being reviewed. The second alternative can be used to answer meta-analytic questions regarding sample characteristics, as well as to calculate the overall magnitude of the correlation.

The final step in meta-analysis involves testing the homogeneity of effect sizes across studies. The variation among study effect sizes can be analyzed using Hedges's Q test of homogeneity. If studies in meta-analysis provide a homogeneous estimate of a combined effect size across studies, then it is more likely that the various studies are testing the same hypothesis. However, if these estimates are heterogeneous, that is, between-study differences are due to unobserved random sources, an effort must be made to identify sample and study characteristics that explain the difference across the studies through the coding process. When combining the outcomes from different studies, one may also choose to use a fixed or random-effects model. The *fixed-effect model* assumes that studies in the meta-analysis use identical methods, whereas the *random-effects model* assumes that studies are a random sample from the universe of all possible studies. The former model considers within-study variability as the only source of variation, while the latter model considers both within-study and between-study variation as sources of differences. Fixed and random-effects models can yield very different results because fixed-effect models are likely to underestimate and random-effect models are likely to overestimate error variance when their assumptions are violated.

Thus, meta-analysis, like any other survey research undertaking, is an observational study of evidence. It has its own limitations and therefore should be undertaken rigorously by using well-defined criteria for selecting and coding individual studies, estimating effect size, aggregating significance levels, and integrating effects.

*Selcuk R. Sirin*

***See also*** Cohen's *d* Statistic; Effect Size, Measures of; Fixed-Effects Models; Homogeneity of Variance; Inclusion Criteria; "Meta-Analysis of Psychotherapy Outcome Studies"; Mixed- and Random-Effects Models; Odds Ratio; Random-Effects Models

**Further Readings**

American Psychological Association. (2009). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Cook, D. J., Guyatt, G. H., Ryan, G., Clifton. J., Buckingham, L., Willan, A., et al. (1993). Should unpublished data be included in meta-analyses? Current convictions and controversies. *JAMA, 269,* 2749–2753.

Cook, T. D., & Leviton, L. C. (1980). Reviewing the literature: A comparison of traditional methods with meta-analysis. *Journal of Personality, 48,* 449–472.

Cooper, H. M. (1989). *Integrating research: A guide for literature reviews* (2nd ed.). Newbury Park, CA: Sage.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.
Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. New York: Academic Press.
Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University Press.
Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.
Rosenthal, R. (1991). *Meta-analytic procedures for social research* (Rev. ed.). Newbury Park, CA: Sage.
Stock, W. A. (1994). Systematic coding for research synthesis. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 125–138). New York: Russell Sage.

# "Meta-Analysis of Psychotherapy Outcome Studies"

The article "Meta-Analysis of Psychotherapy Outcome Studies," written by Mary Lee Smith and Gene Glass and published in *American Psychologist* in 1977, initiated the use of meta-analysis as a statistical tool capable of summarizing the results of numerous studies addressing a single topic. In meta-analysis, individual research studies are identified according to established criteria and treated as a population, with results from each study subjected to coding and entered into a database, where they are statistically analyzed. Smith and Glass pioneered the application of meta-analysis in research related to psychological treatment and education. Their work is considered a major contribution to the scientific literature on psychotherapy and has spurred hundreds of other meta-analytic studies since its publication.

## Historical Context

Smith and Glass conducted their research both in response to the lingering criticisms of psychotherapy lodged by Hans Eysenck beginning in 1952 and in an effort to integrate the increasing volume of studies addressing the efficacy of psychological treatment. In a scathing review of psychotherapy, Eysenck had asserted that any benefits derived from treatment could be attributed to the spontaneous remission of psychological symptoms rather than to the therapy applied. His charge prompted numerous studies on the efficacy of treatment, often resulting in variable and conflicting findings.

Prior to the Smith and Glass article, behavioral researchers were forced to rely on a narrative synthesis of results or on an imprecise tallying method to compare outcome studies. Researchers from various theoretical perspectives highlighted studies that supported their work and dismissed or disregarded findings that countered their position. With the addition of meta-analysis to the repertoire of evaluation tools, however, researchers were able to objectively evaluate and refine their understanding of the effects of psychotherapy and other behavioral interventions. Smith and Glass determined that, on average, an individual who had participated in psychotherapy was better off than 75% of those who were not treated. Reanalyses of the Smith and Glass data, as well as more recent meta-analytic studies, have yielded similar results.

## Effect Size

Reviewing 375 studies on the efficacy of psychotherapy, Smith and Glass calculated an index of effect size to determine the impact of treatment on patients who received psychotherapy versus those assigned to a control group. The effect size was equal to the difference between the means of the experimental and control groups divided by the standard deviation of the control group. A positive effect size communicated the efficacy of a psychological treatment in standard deviation units. Smith and Glass found an effect size of .68, indicating that after psychological treatment, individuals who had completed therapy were superior to controls by .68 standard deviations, an effect size that is generally classified as moderately large.

## Other Findings

While best known for its contribution to research on the general efficacy of psychotherapy, the Smith and Glass study also examined relative efficacy of specific approaches to therapy by classifying studies into 10 theoretical types and calculating an effect size for each. Results indicated that approximately 10% of the variance in the effects of

treatment could be attributed to the type of therapy employed, although the results were confounded by differences in the individual studies, including the number of variables, the duration of treatment, the severity of the presenting problem, and the means by which progress was evaluated. The authors attempted to address these problems by collapsing the 10 types of therapies into four classes: ego therapies, dynamic therapies, behavioral therapies, and humanistic therapies, and then further collapsing the types of therapy into two superclasses labeled *behavioral* and *nonbehavioral* therapies. They concluded that differences among the various types of therapy were negligible. They also asserted that therapists' degrees and credentials were unrelated to the efficacy of treatment, as was the length of therapy.

### Criticisms

Publication of the Smith and Glass article prompted a flurry of responses from critics, including Eysenck, who argued that the studies included in the meta-analysis were too heterogeneous to be compared and that many were poorly designed. Some critics pointed out that an unspecified proportion of studies included in the analysis did not feature an untreated control group. Further, some studies did not have a placebo control group to rule out the effects of attention or expectation among patients. A later reanalysis of the data by Janet Landman and Robyn Dawes published in 1982 used more stringent criteria and featured separate analyses that used only studies that included placebo controls. Their analyses reached conclusions that paralleled those of Smith and Glass.

### Influence

The Smith and Glass study not only altered the landscape of the psychotherapeutic efficacy battle; it also laid the groundwork for meta-analytic studies investigating a variety of psychological and educational interventions. Their work provided an objective means of determining the outcome of a given intervention, summarizing the results of large numbers of studies, and indicating not only whether a treatment makes a difference, but how much of a difference.

At the time of the Smith and Glass publication, the statistical theory of meta-analysis was not yet fully articulated. More recent studies using meta-analysis have addressed the technical problems found in earlier work. As a result, meta-analysis has become an increasingly influential technique in measuring treatment efficacy.

*Sarah L. Hastings*

***See also*** Control Group; Effect Size, Measures of; Meta-Analysis

#### Further Readings

Chambliss, C. H. (2000). A review of relevant psychotherapy outcome research. In C. H. Chambliss (Ed.), *Psychotherapy and managed care: Reconciling research and reality* (pp. 197–214). Boston: Allyn and Bacon.

Eysenck, H. J. (1952). The effects of psychotherapy. *Journal of Consulting Psychology, 16,* 319–324.

Landman, J. T., & Dawes, R. M. (1982). Psychotherapy outcome: Smith and Glass' conclusions stand up under scrutiny. *American Psychologist, 37*(5), 504–516.

Lipsey, M., & Wilson, D. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48*(12), 1181–1209.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32,* 752–760.

Wampold, B. E. (2000). Outcomes of individual counseling and psychotherapy: Empirical evidence addressing two fundamental questions. In S. D. Brown & R. W. Lent (Eds.), *Handbook of counseling psychology* (3rd ed.). New York: Wiley.

## Methods Section

The purpose of a methods section of a research paper is to provide the information by which a study's validity is judged. It must contain enough information so that (a) the study could be repeated by others to evaluate whether the results are reproducible, and (b) others can judge whether the results and conclusions are valid. Therefore, the methods section should provide a clear and precise description of how a study was done and the rationale for the specific procedures chosen.

Historically, the methods section was referred to as the "materials and methods section" to emphasize the two areas that must be addressed. "Materials" referred to what was studied (e.g., humans, animals, tissue cultures), treatments applied, and instruments used. "Methods" referred to the selection of study subjects, data collection, and data analysis. In some fields of study, because "materials" does not apply, alternative headings such as "subjects and methods," "patients and methods," or simply "methods" have been used or recommended.

Below are the items that should be included in a methods section.

## Subjects or Participants

If human or animal subjects were used in the study, who the subjects were and how they were relevant to the research question should be described. Any details that are relevant to the study should be included. For humans, these details include gender, age, ethnicity, socioeconomic status, and so forth, when appropriate. For animals, these details include gender, age, strain, weight, and so forth. The researcher should also describe how many subjects and how they were selected. The selection criteria and rationale for enrolling subjects into the study must be stated explicitly. For example, the researcher should define study and comparison subjects and the inclusion and exclusion criteria of subjects. If the subjects were human, the type of reward or motivation used to encourage them to participate should be stated. When working with human or animal subjects, there must be a declaration that an ethics or institutional review board has determined that the study protocol adheres to ethical principles. In studies involving animals, the preparations made prior to the beginning of the study must be specified (e.g., use of sedation and anesthesia).

## Study Design

The design specifies the sequence of manipulations and measurement procedures that make up the study. Some common designs are experiments (e.g., randomized trials, quasi-experiments), observational studies (e.g., prospective or retrospective cohort, case–control, cross-sectional), qualitative methods (e.g., ethnography, focus groups) and others (e.g., secondary data analysis, literature review, meta-analysis, mathematical derivations, and opinion–editorial pieces). Here is a brief description of the designs. Randomized trials involve the random allocation by the investigator of subjects to different interventions (treatments or conditions). Quasi-experiments involve nonrandom allocation. Both cohort (groups based on exposures) and case–control (groups based on outcomes) studies are longitudinal studies in which exposures and outcomes are measured at different times. Cross-sectional studies measure exposures and outcomes at a single time. Ethnography uses fieldwork to provide a descriptive study of human societies. A focus group is a form of qualitative research in which people assembled in a group are asked about their attitude toward a product or concept. An example of secondary data is the abstraction of data from existing administrative databases. A meta-analysis combines the results of several studies that address a set of related research hypotheses.

## Data Collection

The next step in the methods section is a description of the variables that were measured and how these measurements were made. In laboratory and experimental studies, the description of measurement instruments and reagents should include the manufacturer and model, calibration process, and how measurements were made. In epidemiologic and social studies, the development and pretest of questionnaires, training of interviewers, data extraction from databases, and conduct of focus groups should be described where appropriate. In some cases, the survey instrument (questionnaire) may be included as an appendix to the research paper.

## Data Analysis

The last step in the methods section is to describe the way in which the data will be presented in the results section. For quantitative data, this step should specify whether and which

statistical tests will be used for making the inference. If statistical tests are used, this part of the methods section must specify the significance level and whether one- or two sided or the type of confidence intervals. For qualitative data a common analysis is observer impression. That is, expert or lay observers examine the data, form an impression, and report their impression in a structured, quantitative form.

The following are some tips for writing the methods section: (a) The writing should be direct and precise. Complex sentence structures and unimportant details should be avoided. (b) The rationale or assumptions on which the methods are based may not always be obvious to the audience and so should be explained clearly. This is particularly true when one is writing for a general audience, as opposed to a subspecialty group. The writer must always keep in mind who the audience is. (c) The methods section should be written in the past tense. (d) Subheadings, such as participants, design, and so forth, may help readers navigate the paper. (e) If the study design is complex, it may be helpful to include a diagram, table, or flowchart to explain the methods used. (f) Results should not be placed in the methods section. However, the researchers may include preliminary results from a pilot test they used to design the main study they are reporting.

The methods section is important because it provides the information the reader needs to judge the study's validity. It should provide a clear and precise description of how a study was conducted and the rationale for specific study methods and procedures.

*Bernard Choi and Anita Pak*

***See also*** Discussion Section; Results Section; Validity of Research Conclusions

### Further Readings

Branson, R. D. (2004). Anatomy of a research paper. *Respiratory Care, 49,* 1222–1228.

Hulley, S. B., Newman, T. B., & Cummings, S. R. (1988). The anatomy and physiology of research. In S. B. Hulley & S. R. Cummings (Eds.), *Designing clinical research* (pp. 1–11). Baltimore: William & Wilkins.

Kallet, R. H. (2004). How to write the methods section of a research paper. *Respiratory Care, 49,* 1229–1232.

Van Damme, H., Michel, L., Ceelen, W., & Malaise, J. (2007). Twelve steps to writing an effective "materials and methods" section. *Acta Chirurgica Belgica, 107,* 102.

## Method Variance

Method is what is used in the process of measuring something, and it is a property of the measuring instrument. The term *method effects* refers to the systematic biases caused by the measuring instrument. *Method variance* refers to the amount of variance attributable to the methods that are used. In psychological measures, method variance is often defined in relationship to trait variance. *Trait variance* is the variability in responses due to the underlying attribute that one is measuring. In contrast, method variance is defined as the variability in responses due to characteristics of the measuring instrument. After sketching a short history of method variance, this entry discusses features of measures and method variance analyses and describes approaches for reducing method effects.

### A Short History

No measuring instrument is free from error. This is particularly germane in social science research, which relies heavily on self-report instruments. Donald Thomas Campbell was the first to mention the problem of method variance. In 1959, Campbell and Donald W. Fiske described the fallibility inherent in all measures and recommended the use of multiple methods to reduce error. Because no single method can be the gold standard for measurement, they proposed that multiple methods be used to triangulate on the underlying "true" value. The concept was later extended to unobtrusive measures.

Method variance has not been well defined in the literature. The assumption has been that the reader knows what is meant by *method variance*. It is often described in a roundabout way, in relationship to trait variance. Campbell and Fiske pointed out that there is no fixed demarcation between trait and method. Depending on the goals

of a particular research project, a characteristic may be considered either a method or a trait. Researchers have reported the methods that they use as different tests, questionnaires with different types of answers, self-report and peer ratings, clinician reports, or institutional records, to name a few.

In 1950 Campbell differentiated between structured and nonstructured measures, along with those whose intent was disguised, versus measures that were obvious to the test taker. Later Campbell and others described the characteristics associated with unobtrusive methods, such as physical traces and archival records. More recently, Lee Sechrest and colleagues extended this characterization to observable methods.

Others have approached the problem of method from an "itemetric" level, in paper-and-pencil questionnaires. A. Angleitner, O. P. John, and F. Löhr proposed a series of item-level characteristics, including overt reactions, covert reactions, bodily symptoms, wishes and interests, attributes of traits, attitudes and beliefs, biographical facts, others' reactions, and bizarre items.

## Obvious Methods

There appear to be obvious, or manifest, features of measurement, and these include stimulus formats, response formats, response categories, raters, direct rating versus summative scale, whether the stimulus or response is rated, and finally, opaque versus transparent measures. These method characteristics are usually mentioned in articles to describe the methods used. For example, an abs-tract may describe a measure as "a 30-item true–false test with three subscales," "a structured interview used to collect school characteristics," or "patient functioning assessed by clinicians using a 5-point scale."

### *Stimulus and Response Formats*

The stimulus format is the ways the measure is presented to the participant, such as written or oral stimulus. The response format refers to the methods used to collect the participant's response and includes written, oral, and graphical approaches. The vast majority of stimulus and response formats used in social science research are written paper-and-pencil tests.

### *Response Categories*

The response categories include the ways an item may be answered. Examples of response categories include multiple-choice items, matching, Likert-type scales, true–false answers, responses to open-ended questions, and visual analogue scales. Close-ended questions are used most frequently, probably because of their ease of administration and scoring. Open-ended questions are used less frequently in social science research. Often the responses to these questions are very short, or the question is left blank. Open-ended questions require extra effort to code. Graphical responses such as visual analogue scales are infrequently used.

### *Raters*

Raters are a salient method characteristic. Self-report instruments comprise the majority of measures. In addition to the self as rater, other raters include, for example, teachers, parents, and peers. Other raters may be used in settings with easy access to them. For example, studies conducted in schools often include teacher ratings and may collect peer and parent ratings. Investigations in medical settings may include ratings by clinicians and nurses.

The observability of the trait in question probably determines the accuracy of the ratings by others. An easily observable trait such as extroversion will probably generate valid ratings. However, characteristics that cannot be seen, particularly those that the respondent chooses to hide, will be harder to rate. Racism is a good example of a characteristic that may not be amenable to ratings.

### *Direct Versus Summative Scale*

This method characteristic refers to the number of items used to measure a characteristic. The respondent may be asked directly about his or her standing on a trait; for example, How extroverted are you? In other instances, multiple-item scales are employed. The items are then summed to estimate the respondent's standing on the trait. Direct,

single items may be sufficient if a trait is obvious and/or the respondent does not care about the results.

### *Rating the Stimulus Versus Rating the Response*

Rating the prestige of colleges or occupations is an example of rating the stimulus; self-report questionnaires for extroversion or conscientiousness are examples of rating the response. The choice depends on the goals of the study.

### *Opaque Versus Transparent Measures*

This method characteristic refers to whether the purpose of a test is easily discerned by the respondent. The Stanford-Binet is obviously a test of intelligence, and the Myers-Briggs Type Indicator inventory measures extroversion. These are transparent tests. If the respondent cannot easily guess the purpose of a test, it is opaque.

## Types of Analyses Used for Method Variance

If a single method is used, it is not possible to estimate method effects. Multiple methods are required in an investigation in order to study method effects. When multiple methods are collected, they must be combined in some way to estimate the underlying trait. Composite scores or latent factor models are used to estimate the trait. If the measures in a study have used different sources of error, the resulting trait estimate will contain less method bias.

Estimating the effect of methods is more complicated. Neal Schmitt and Daniel Stutts have provided an excellent summary of the types of analyses that may be used to study method variance. Currently, the most popular method of analysis for multitrait–multimethod matrices is *confirmatory factor analysis*. However, there are a variety of problems inherent in this method, and *generalizability theory analysis* shows promise for multitrait–multimethod data.

### *Does Method Variance Pose a Real Problem?*

The extent of variance attributable to methods has not been well studied, although several interesting articles have focused on it. Joseph A. Cote and M. Ronald Buckley examined 70 published studies and reported that trait accounted for more than 40% of the variance and method accounted for approximately 25%. D. Harold Doty and William H. Glick obtained similar results.

## Reducing Effects of Methods

A variety of approaches can be used to lessen the effects of methods in research studies. Awareness of the problem is an important first step. The second is to avoid measurement techniques laden with method variance. Third, incorporate multiple measures that use maximally different methods, with different sources of error variance. Finally, the multiple measures can be combined into a trait estimate during analysis. Each course of action reduces the effects of methods in research studies.

*Melinda Fritchoff Davis*

***See also*** Bias; Confirmatory Factor Analysis; Construct Validity; Generalizability Theory; Multitrait–Multimethod Matrix; Rating; Triangulation; True Score; Validity of Measurement

### Further Readings

Campbell, D. T. (1950). The indirect assessment of social attitudes. *Psychological Bulletin, 47,* 15–38.

Cote, J. A., & Buckley, M. R. (1987). Estimating trait, method, and error variance: Generalizing across 70 construct validation studies. *Journal of Marketing Research, 24,* 315–318.

Doty, D. H., & Glick, W. H. (1998). Common methods bias: Does common methods variance really bias results? *Organizational Research Methods, 1*(4), 374–406.

Schmitt, N., & Stutts, D. M. (1986). Methodology review: Analysis of multitrait-multimethod matrices. *Applied Psychological Measurement, 10,* 1–22.

Sechrest, L. (1975). Another look at unobtrusive measures: An alternative to what? In W. Sinaiko & L. Broedling (Eds.), *Perspectives on attitude assessment: surveys and their alternatives* (pp. 103–116). Washington, DC: Smithsonian Institution.

Sechrest, L., Davis, M. F., Stickle, T., & McKnight, P. (2000). Understanding "method" variance. In L. Bickman (Ed.), *Research Design: Donald Campbell's Legacy* (pp. 63–88). Thousand Oaks, CA: Sage.

Webb, E. T., Campbell, D. T., Schwartz, R. D., Sechrest, L., & Grove, J. B. (1981). *Nonreactive measures in*

*the social sciences* (2nd ed.). Boston: Houghton Mifflin.

# Missing Data, Imputation of

Imputation involves replacing missing values, or *missings*, with an estimated value. In a sense, imputation is a prediction solution. It is one of three options for handling missing data. The general principle is to delete when the data are expendable, impute when the data are precious, and segment for the less common situation in which a large data set has a large fissure. Imputation is measured against deletion; it is advantageous when it affords the more accurate data analysis of the two. This entry discusses the differences between imputing and deleting, the types of missings, the criteria for preferring imputation, and various imputation techniques. It closes with application suggestions.

## Impute or Delete

The trade-off is between inconvenience and bias. There are two choices for deletion (casewise or pairwise) and several approaches to imputation. *Casewise deletion* omits entire observations (or cases) with a missing value from all calculations. *Pairwise deletion* omits observations on a variable-by-variable basis. Casewise deletion sacrifices partial information either for convenience or to accommodate certain statistical techniques. Techniques such as structural equation modeling may require complete data for all the variables, so only casewise deletion is possible for them. For techniques such as calculating correlation coefficients, pairwise deletion will leverage the partial information of the observations, which can be advantageous when one is working with small sample sizes and when missings are not random.

Imputation is the more advantageous technique when (a) the missings are not random, (b) the missings represent a large proportion of the data set, or (c) the data set is small or otherwise

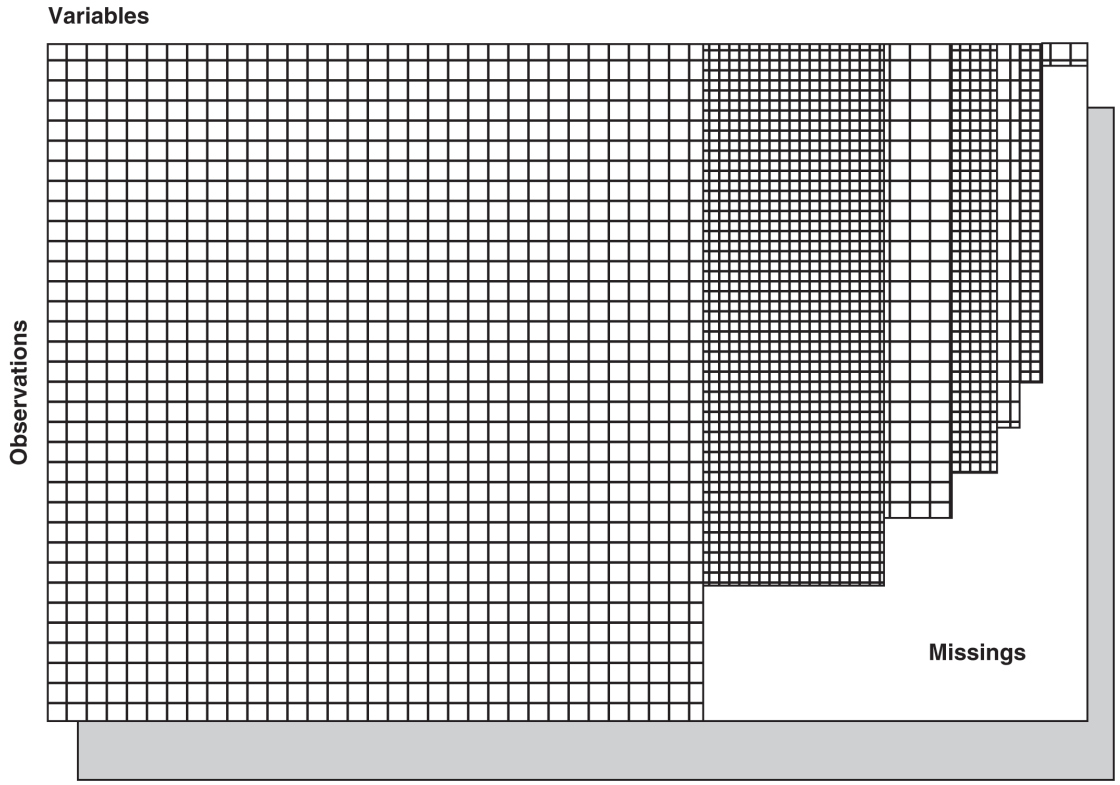



**Figure 1** Missing Data Structure

precious. If the missings do not occur at random, which is the most common situation, then deleting can create significant bias. For some situations, it is possible to repair the bias through weighting—as in poststratification for surveys. If the data set is small or otherwise precious, then deleting can severely reduce the statistical power or value of the data analysis.

Imputation can repair the missing data by creating one or more versions of how the data set should appear. By leveraging external knowledge, good technique, or both, it is possible to reduce bias due to missing values. Some techniques offer a quick improvement over deletion. Software is making these techniques faster and sharper; however, the techniques should be conducted by those with appropriate training.

## Categorizing Missingness

Missingness can be categorized in two ways: the physical structure of the missings and the underlying nature of the missingness. First, the structure of the missings can be due to item or unit missingness, the merging of structurally different data sets, or barriers attributable to the data collection tools. Item missingness refers to the situation in which a single value is missing for a particular observation, and unit missingness refers to the situation in which all the values for an observation are missing. Figure 1 provides an illustration of missingness.

Second, missings can be categorized by the underlying nature of the missingness. These three categories are (1) missing completely at random (MCAR), (2) missing at random (MAR), and (3) missing not at random (MNAR), summarized in Table 1 and discussed below.

Categorizing missings into one of these three groups provides better judgment as to the most appropriate imputation technique and the ramifications of employing that technique. MCAR is the least common, yet the easiest to address. MAR can be thought of as missing partially at random; the point is that there is some pattern that can be leveraged. There are statistical tests for inferring MCAR and MAR. There are many imputation techniques geared toward MAR. The potential of these techniques depends on the degree to which other variables are related to the missings. MNAR is also known as *informative missing, nonignorable missingness*. It is the most difficult to address. The most promising approach is to use external data to identify and repair this problem.

**Table 1** Underlying Nature of Missingness

| *Type* | *Definitions and Examples* | *Most Likely Approach* |
|---|---|---|
| Missing completely at random | Missing values occur completely at random with no relationship to themselves or the values of any other variables. For example, suppose researchers can only afford to measure temperature for a randomly selected subsample. The remaining observations will be missing, completely at random. | Delete |
| Missing at random or missing partially at random | Missing values occur partially at random, occurring relative to observable variables and otherwise randomly. That is, after controlling for all other observable variables, the missing values are random. | Impute |
| Missing not at random; informative missing | Missing values occur relative to the variable itself and are not random after controlling for all other variables. For example, as temperatures become colder, the thermometer is increasingly likely to fail and other variables cannot fully explain the pattern in these failures. | Impute |

## Statistical and Contextual Diagnosis

Analyzing the missings and understanding their context can help researchers infer whether they are MCAR, MAR, or MNAR. There are five considerations: (1) relationships between the missings and other variables, (2) concurring missing patterns, (3) relationships between variables and external information, (4) context of the analysis, and (5) software.

### *Relationships Between the Missings and Other Variables*

Exploratory data analysis will reveal relationships, which can indicate MAR and lead to an appropriate imputation technique. There are statistical tests for comparing the means and distributions of covariates for missings versus nonmissings. Statistically significant results imply MAR and therefore discount MCAR. A number of other techniques are available to help provide insight, including logistic regression, regression trees, and cluster analysis. The insight from this data analysis should be juxtaposed with the context of the application and the source of the data.

### *Concurring Missing Patterns*

Concurring missings suggest a shared event. Placed within context, these can indicate that the missings were "manufactured" by the data collection tool or the data pipeline. These clues can suggest MCAR, MAR, or MNAR.

### *Relationships Between Variables and External Information*

Comparing the distribution of variables with missing values to external data can reveal that the data are MNAR rather than MCAR. For example, the mean of the nonmissings should be compared with an external estimate of the overall mean.

### *Context of the Analysis*

It is necessary to study the context of the analysis, including the statistical aspects of the problem and the consequences of various imputation techniques. For example, missings belonging to controlled variables in an experiment have different implications from those of observed variables.

Missing values do not always represent a failure to measure. Sometimes, these are accurate measurements. For example, if amount of loan loss is missing, it could be because the loss amount was not entered into the computer or because the loan never defaulted. In the latter case, the value can be thought of as either zero or as "does not apply," whichever is more appropriate for the analysis. It is common for collection tools to record a missing value when a zero is more appropriate.

### *Software*

Software has become more than a tool. Investigators' choices are often limited to those techniques supported by available software.

## Imputation Techniques

Imputation techniques are optimal when missing values are not all equal, as in MAR. All the techniques raise statistical issues. Criteria for choosing a technique include the underlying nature of the missings (MCAR, MAR, or MNAR), speed, implications regarding bias and variance estimation, and considerations related to preserving the natural distribution and correlation structure. A final consideration is to avoid extrapolating outside the range space of the data. Some missing values have logical bounds. This entry discusses five types of techniques: (1) substitution, (2) regression (least squares), (3) Bayesian methods, (4) maximum likelihood estimation (MLE)–expectation maximization algorithm, and (5) multiple imputation.

### *Substitution Techniques*

One quick solution is to substitute the mean, median, series mean, a linear interpolation, and so forth, for the missings. One drawback for using the global mean or median is that values are repeated. This can create a spike in the distribution. It can be avoided by substituting a local mean or median. These quick substitutions tend to result in underestimating the variance and can inflate or deflate correlations.

For categorical variables, characteristic analysis compares the mean of the missings to the means of

the other categories. The missings are assigned the category with the closest mean.

*Hot deck* and *cold deck* are techniques for imputing real data into the missings, with or without replacement. For hot deck, the donor data are the same data set, and for cold deck, the donor data are another data set. Hot deck avoids extrapolating outside the range space of the data set, and it better preserves the natural distribution than does imputation of a mean. Both tend to be better for MAR.

### *Regression (Least Squares)*

Regression-based imputation predicts the missings on the basis of ordinary–least-squares or weighted–least-squares modeling of the nonmissing data. This assumes that relationships among the nonmissing data extrapolate to the missing-value space. This technique assumes that the data are MAR and not MCAR. It creates bias depending on the degree to which the model is overfit. As always, validation techniques such as *bootstrapping* or *data splitting* will curb the amount of overfitting.

Regression-based imputation underestimates the variance. Statisticians have studied the addition of random errors to the imputed values as a technique to correct this underestimation. The random errors can come from a designated distribution or from the observed data.

Regression-based imputation does not preserve the natural distribution or respect the associations between variables. Also, it repeats imputed values when the independent variables are identical.

### *Bayesian Methods*

The approximate Bayesian bootstrap uses logistic regression to predict missing and nonmissing values for the dependent variable, *y,* based on the observed *x* values. The observations are then grouped on the basis of the probability of the value missing. Candidate imputation values are randomly selected, with replacement, from the same group.

### *MLE–Expectation Maximization Algorithm*

The expectation maximization algorithm is an iterative, two-step approach for finding an MLE for imputation. The initial step consists of deriving an expectation based on latent variables. This is followed by a maximization step, computing the MLE. The technique assumes an underlying distribution, such as the normal, mixed normal, or Student's *t*.

The MLE method assumes that missing values are MAR (as opposed to MCAR) and shares with regression the problem of overfitting. MLE is considered to be stronger than regression and to make fewer assumptions.

### *Multiple Imputation*

Multiple imputation leverages another imputation technique to impute and reimpute the missings. This technique creates multiple versions of the data set; analyzes each one; and then combines the results, usually by averaging. The advantages are that this process is easier than MLE, robust to departures from underlying assumptions, and provides better estimates of variance than regression does.

## Suggestions for Applications

A project's final results should include reasons for deleting or imputing. It should justify any selected imputation technique and enumerate the corresponding potential biases. As a check, it is advisable to compare the results obtained with imputations and those obtained without them. This comparison will reveal the effect due to imputation. Finally, there is an opportunity to clarify wheather the missings provide an additional hurdle or valuable information.

*Randy Bartlett*

***See also*** Bias; Data Cleaning; Outlier; Residuals

### Further Readings

Efron, B. (1994). Missing data, imputation, and the bootstrap. *Journal of the American Statistical Association, 89*(426), 463–475.

Little, R. J. A. (1992). Regression with missing x's: A review. *Journal of the American Statistical Association, 87*(420), 1227–1237.

Little, R. J. A., & D. B. Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.

Schaefer, J., & Graham, J. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147–177.

# Mixed- and Random-Effects Models

Data that are collected or generated in the context of any practical problem always exhibit variability. This variability calls for the use of appropriate statistical methodology for the data analysis. Data that are obtained from designed experiments are typically analyzed using a model that takes into consideration the various sources or *factors* that could account for the variability in the data. Here the term *experiment* denotes the process by which data are generated based on the basis of planned changes in one or more input variables that are expected to influence the response. The plan or layout used to carry out the experiment is referred to as an *experimental design* or *design of the experiment*. The analysis of the data is based on an appropriate statistical model that accommodates the various factors that explain the variability in the data. If all the factors are fixed, that is, nonrandom, the model is referred to as a *fixed-effects model*. If all the factors are random, the model is referred to as a *random-effects model*. On the other hand, if the experiment involves fixed as well as random factors, the model is referred to as a *mixed-effects model*. In this entry, mixed- as well as random-effects models are introduced through some simple research design examples. Data analysis based on such models is briefly commented on.

## A Simple Random-Effects Model

Here is a simple example, taken from Douglas C. Montgomery's book on experimental designs. A manufacturer wishes to investigate the research question of whether batches of raw materials furnished by a supplier differ significantly in their calcium content. Suppose data on the calcium content will be obtained on five batches that were received in 1 day. Furthermore, suppose six determinations of the calcium content will be made on each batch. Here *batch* is an input variable, and the response is the calcium content. The input variable is also called a *factor*. This is an example of a single-factor experiment, the factor being *batch*. The different possible categories of the factor are referred to as *levels* of the factor. Thus the factor *batch* has five levels. Note that on each level of the factor (i.e., on each batch), we obtain the same number of observations, namely six. In such a case, the data are called *balanced*. In some applications, it can happen that the numbers of observations obtained on each level of the factor are not the same; that is, we have unbalanced data.

In the above example, suppose the purpose of the data analysis is to test whether there is any difference in the calcium content among five given batches of raw materials obtained in one day. In experimental design terminology, we want to test whether the five batches have the same *effects*. This example, as stated, involves a factor (namely batch) having *fixed effects*. The reason is that there is nothing random about the batches themselves; the manufacturer has five batches given to him on a single day, and he wishes to make a comparison of the calcium content among the five given batches. However, there are many practical problems in which the factor could have random effects. In the context of the same example, suppose a large number of batches of raw materials are available in the warehouse, and the manufacturer does not have the resources to obtain data on all the batches regarding their calcium content. A natural option in this case is to collect data on a sample of batches, randomly selected from the population of available batches. Random selection is done to ensure that we have a representative sample of batches. Note that if another random selection is made, a different set of five batches could have been selected. If five batches are selected randomly, we then have a factor having *random effects*. Note that the purpose of our data analysis is not to draw conclusions regarding the calcium content of the five batches randomly selected; rather, we would like to use the random sample of five batches to draw conclusions regarding the population of all batches. The difference between fixed effects and random effects should now be clear. In the fixed-effects case, we have a given number of levels of a factor, and the purpose of the data analysis is to make comparisons among these given levels only, based on the responses that have been obtained. In the random-effects case, we make a random selection of a few levels of the factor (from a population of levels), and the responses are obtained on the randomly selected levels only. However, the purpose of the analysis is to make inferences concerning the population of all levels.

In order to make the concepts more concrete, let $y_{ij}$ denote the $j$th response obtained on the $i$th level of the factor, where $j = 1, 2, \ldots, n$, and $i = 1, 2, \ldots, a$. Here $a$ denotes the number of levels of the factor, and $n$ denotes the number of responses obtained on each level. For our example, $a = 5$, $n = 6$, and $y_{ij}$ is the $j$th determination of the calcium content from the $i$th batch of raw material. The data analysis can be done assuming the following structure for the $y_{ij}$s, referred to as a model:

$$y_{ij} = \mu + \tau_i + e_{ij}, \tag{1}$$

where $\mu$ is a common mean, the quantity $\tau_i$ represents the effect due to the $i$th level of the factor (effect due to the $i$th batch), and $e_{ij}$ represents experimental error. The $e_{ij}$s are assumed to be random, following a normal distribution with mean zero and variance $\sigma^2$. In the fixed-effects case, the $\tau_i$s are fixed unknown parameters, and the problem of interest is to test whether the $\tau_i$s are equal. The model for the $y_{ij}$ is now referred to as a *fixed-effects model*. In this case, the restriction $\sum_{i=1}^{a} \tau_i = 0$ can be assumed, without loss of generality. In the random-effects case, the $\tau_i$s are assumed to be random variables following a normal distribution with mean zero and variance $\sigma_\tau^2$. The model for the $y_{ij}$ is now referred to as a *random-effects model*. Note that $\sigma_\tau^2$ is a population variance; that is, it represents the variability among the population of levels of the factor. Now the problem of interest is to test the hypothesis that $\sigma_\tau^2$ is zero. If this hypothesis is accepted, then the conclusion is that the different levels of the factor do not exhibit significant variability among them. In the context of the example, if the batches are randomly selected, and if the hypothesis $\sigma_\tau^2 = 0$ is not rejected, then the data support the conclusion that there is no significant variability among the different batches in the population.

## Mixed- and Random-Effects Models for Multifactor Experiments

In the context of the same example, suppose the six calcium content measurements on each batch are made by six different operators. While carrying out the measuring process, there could be differences among the operators. In other words, in addition to the effect due to the batches, there exist effects due to the operators as well, accounting for the differences among them. A possible model that could capture both the batch effects and the operator effects is

$$y_{ij} = \mu + \tau_i + \beta_j + e_{ij}, \tag{2}$$

$i = 1, 2, \ldots, a$, $j = 1, 2, \ldots, b$, where $y_{ij}$ is the calcium content measurement obtained from the $i$th batch by the $j$th operator; $\beta_j$ is the effect due to the $j$th operator; and $\mu$, the $\tau_i$s, and the $e_{ij}$s are as defined before. In the context of the example, $a = 5$ and $b = 6$. Now there are two input variables, that is, batches and operators, that are expected to influence the response (i.e., the calcium content measurement). This is an example of a two-factor experiment. Note that if the batches as well as the operators are randomly selected, then the $\tau_i$s, as well as the $\beta_j$s, become random variables; the above model is then called a random-effects model. However, if only the batches are randomly selected (so that the $\tau_i$s are random), but the measurements are taken by a given group of operators (so that the $\beta_j$s are fixed unknown parameters), then we have a mixed-effects model. That is, the model involves fixed effects corresponding to the given levels of one factor and random-effects corresponding to a second factor, whose levels are randomly selected. When the $\beta_j$s are fixed, the restriction $\sum_{j=1}^{b} \beta_j = 0$ may be assumed. For a random-effects model, independent normal distributions are typically assumed for the $\tau_i$s, for the $\beta_j$s, and the $e_{ij}$s, similar to that for Model 1. When the effects due to a factor are random, the hypothesis of interest is whether the corresponding variance is zero. In the fixed-effects case, we test whether the effects are the same for the different levels of the factor.

Note that Model 2 makes a rather strong assumption, namely, that the combined effect due to the two factors, batch and operator, can be written as the sum of an effect due to the batch and an effect due to the operator. In other words, there is *no interaction* between the two factors. In practice, such an assumption may not always hold when responses are obtained based on the combined effects of two or more factors. If interaction is present, the model should include the combined effect due to the two factors. However, now multiple measurements are necessary to carry out the data

analysis. Thus, suppose each operator makes three determinations of the calcium content on each batch. Denote by $y_{ijk}$ the $k$th calcium content measurement on the $i$th batch by the $j$th operator. When there is interaction, the assumed model is

$$y_{ijk} = \mu + \tau_i + \beta_j + \gamma_{ij} + e_{ijk}, \qquad (3)$$

where $i = 1, 2, \ldots, a$; $j = 1, 2, \ldots, b$; and $k = 1, 2, \ldots, n$ (say). Now $\tau_i$ represents an average effect due to the $i$th batch. In other words, consider the combined effect due to the $i$th batch and the $j$th operator, and average it over $j = 1, 2, \ldots, b$. We refer to $\tau_i$ as the *main effect* due to the $i$th batch. Similarly, $\beta_j$ is the main effect due to the $j$th operator. The quantity $\gamma_{ij}$ is the interaction between the $i$th batch and the $j$th operator. If a set of given batches and operators is available for the experiment (i.e., there is no random selection), then the $\tau_i$s, $\beta_j$s, and $\gamma_{ij}$s are all fixed, and Equation 3 is then a fixed-effects model. On the other hand, if the batches are randomly selected, whereas the operators are given, then the $\tau_i$s and $\gamma_{ij}$s are random, but the $\beta_j$s are fixed. Thus Model 3 now becomes a mixed-effects model. However, if the batches and operators are both randomly selected, then all the effects in Equation 3 are random, resulting in a random-effects model. Equation 3 is referred to as a *two-way classification model with interaction*. If the batches are randomly selected, whereas the operators are given, then a formal derivation of Model 3 will result in the conditions $\sum_{j=1}^{b} \beta_j = 0$ and $\sum_{j=1}^{b} \gamma_{ij} = 0$ for every $i$. That is, the random variables $\gamma_{ij}$s satisfy a restriction. In view of this, the $\gamma_{ij}$s corresponding to a fixed $i$ will not be independent among themselves. The normality assumptions are typically made on all the random quantities.

In the context of Model 1, note that two observations from the same batch are correlated. In fact $\sigma_\tau^2$ is also the covariance between two observations from the same batch. Other examples of correlated data where mixed- and random-effects models are appropriate include longitudinal data, clustered data, and repeated measures data. By including random-effects in the model, it is possible for researchers to account for multiple sources of variation. This is indeed the purpose of using mixed- and random-effects models for analyzing data in the physical end engineering sciences, medical and biological sciences, social sciences, and so forth.

## Data Analysis Based on Mixed- and Random-Effects Models

When the same number of observations is obt-ained on the various level combinations of the factors, the data are said to be balanced. For Model 3, balanced data correspond to the situation in which exactly $n$ calcium content measurements (say, $n = 6$) are obtained by each operator from each batch. Thus the observations are $y_{ijk}$; $k = 1, 2, \ldots, n$; $j = 1, 2, \ldots, b$; and $i = 1, 2, \ldots, a$. On the other hand, unbalanced data correspond to the situation in which the number of calcium content determinations obtained by the different operators are not all the same for all the batches. For example, suppose there are five operators, and the first four make six measurements each on the calcium content from each batch, whereas the fifth operator could make only three observations from each batch because of time constraints; we then have unbalanced data. It could also be the case that an operator, say the first operator, makes six calcium content determinations from the first batch but only five each from the remaining batches. For unbalanced data, if $n_{ij}$ denotes the number of calcium content determinations made by the $j$th operator on the $i$th batch, the observations are $y_{ijk}$; $k = 1, 2, \ldots, n_{ij}$; $j = 1, 2, \ldots, b$; and $i = 1, 2, \ldots, a$. The analysis of unbalanced data is considerably more complicated under mixed- and random-effects models, even under normality assumptions. The case of balanced data is somewhat simpler.

### *Analysis of Balanced Data*

Consider the simple Model 1 with balanced data, along with the normality assumptions for the distribution of the $\tau_i$s and the $e_{ij}$s with variances $\sigma_\tau^2$ and $\sigma_e^2$, respectively. The purpose of the data analysis can be to estimate the variances $\sigma_\tau^2$ and $\sigma_e^2$, to test the null hypothesis that $\sigma_\tau^2 = 0$, and to compute a confidence interval for $\sigma_\tau^2$ and sometimes for the ratios $\sigma_\tau^2/\sigma_e^2$ and $\sigma_\tau^2/(\sigma_\tau^2 + \sigma_\tau^2)$. Note that the ratio $\sigma_\tau^2/\sigma_e^2$ provides information on the relative magnitude of $\sigma_\tau^2$ compared with that of $\sigma_e^2$. If the variability in the data is mostly due to the variability among the different levels of the factor, $\sigma_\tau^2$ is expected to be large compared with $\sigma_e^2$, and the hypothesis $\sigma_\tau^2 = 0$ is expected to be rejected. Also note that since the variance of the observations, that

is, the variance of the $y_{ij}$s in Model 1, is simply the sum $\sigma_\tau^2 + \sigma_e^2$, the ratio $\sigma_\tau^2/(\sigma_\tau^2 + \sigma_e^2)$ is the fraction of the total variance that is due to the variability among the different levels of the factor. Thus the individual variances as well as the above ratios have practical meaning and significance.

Now consider Model 3 with random effects and with the normality assumptions $\tau_i \sim N(0, \sigma_\tau^2)$, $\beta_j \sim N(0, \sigma_\beta^2)$, $\gamma_{ij} \sim N(0, \sigma_\gamma^2)$, and $e_{ijk} \sim N(0, \sigma_e^2)$, where all the random variables are assumed to be independently distributed. Now the problems of interest include the estimation of the different variances and testing the hypothesis that the random-effects variances are zeros. For example, if the hypothesis $\sigma_\gamma^2 = 0$ cannot be rejected, we conclude that there is no significant interaction. If Model 3 is a mixed-effects model, then the normality assumptions are made on the effects that are random. Note, however, that the $\gamma_{ij}$s, although random, will no longer be independent in the mixed-effects case, in view of the restriction $\sum_{j=1}^{b} \gamma_{ij} = 0$ for every $i$.

The usual analysis of variance (ANOVA) decomposition can be used to arrive at statistical procedures to address all the above problems. To define the various ANOVA sums of squares for Model 3 in the context of our example on calcium content determination from different batches using different operators, let

$$\bar{y}_{ij.} = \frac{1}{n}\sum_{k=1}^{n} y_{ijk},\ \bar{y}_{i..} = \frac{1}{bn}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk},\ \bar{y}_{.j.} = \frac{1}{an}\sum_{i=1}^{a}\sum_{k=1}^{n} y_{ijk},\ \bar{y}_{...} = \frac{1}{abn}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}.$$

If $SS_\tau$, $SS_\beta$, $SS_\gamma$, and $SS_e$ denote the ANOVA sum of squares due to the batches, operators, interaction, and error, respectively, these are given by

$$SS_\tau = bn\sum_{i=1}^{a}(\bar{y}_{i..} - \bar{y}_{...})^2,\ SS_\beta = an\sum_{j=1}^{b}\left(\bar{y}_{.j.} - \bar{y}_{...}\right)^2$$

$$SS_\gamma = n\sum_{i=1}^{a}\sum_{j=1}^{b}\left(\bar{y}_{ij.} - \bar{y}_{...}\right)^2 - SS_\tau - SS_\beta,$$

$$SS_e = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(y_{ijk} - \bar{y}_{ij.}\right)^2.$$

The following table shows the ANOVA table and the expected mean squares in the mixed-effects case and in the random-effects case; these are available in a number of books dealing with mixed- and random-effects models, in particular in Montgomery's book on experimental designs. In the mixed-effects case, the $\beta_j$s are fixed, but the $\tau_i$s and $\gamma_{ij}$s are random. In other words, the batches are randomly selected, but the operators consist of a fixed group. In the random-effects case, the $\beta_j$s, the $\tau_i$s, and the $\gamma_{ij}$s are all random. In the table, the notations $MS_\tau$ and so forth are used to denote mean squares.

Note that the expected values can be quite different depending on whether we are in the mixed-effects setup or the random-effects setup. Also, when an expected value is a linear combination of only the variances, then the sum of squares, divided by the expected value, has a chi-square distribution. For example, in Table 1, in the mixed-effects case, $SS_\tau/(\sigma_e^2 + bn\sigma_\tau^2)$ follows a chi-square distribution with $a - 1$ degrees of freedom. However, if an expected value also involves the fixed-effects parameters, then the chi-square distribution holds under the appropriate hypothesis concerning the fixed effects. Thus, in Table 1, in the mixed-effects case, $SS_\beta/(\sigma_e^2 + n\sigma_\gamma^2)$ follows a chi-square distribution with $b - 1$ degrees of freedom, under the hypothesis $\beta_1 = \beta_2 = \cdots = \beta_b = 0$. Furthermore, for testing the various hypotheses, the denominator of the $F$ ratio is not always the mean square due to error. If one compares the expected values in Table 1, one can see that for testing $\sigma_\tau^2 = 0$ in the mixed-effects case, the $F$ ratio is $MS_\tau/MS_e$. However, for testing $\beta_1 = \beta_2 = \ldots = \beta_b = 0$ in the mixed-effects case, the $F$ ratio is $MS_\beta/MS_\gamma$. In view of this, it is necessary to know the expected values before we can decide on the appropriate $F$ ratio for testing a hypothesis under mixed- and random-effects models. Fortunately, procedures are available for the easy calculation of the expected values when the data are balanced. The expected values in Table 1 immediately provide us with $F$ ratios for testing all the different hypotheses in the mixed-effects case, as well as in the random-effects case. This is so because, under the hypothesis, we can identify exactly two sums of squares having the same expected value. However, this is

**Table 1** ANOVA and Expected Mean Squares Under Model 3 With Mixed Effects and With Random Effects, When the Data Are Balanced

| *Source of Variability* | *Sum of Squares (SS)* | *Degrees of Freedom* | *Mean Square (MS)* | *Expected Mean Square (Mixed-Effects Case)* | *Expected Mean Square (Random-Effects Case)* |
|---|---|---|---|---|---|
| Batches | $SS_\tau$ | $a-1$ | $MS_\tau$ | $\sigma_e^2 + bn\sigma_\tau^2$ | $\sigma_e^2 + n\sigma_\gamma^2 + bn\sigma_\tau^2$ |
| Operators | $SS_\beta$ | $b-1$ | $MS_\beta$ | $\sigma_e^2 + n\sigma_\gamma^2 + an\frac{\sum_{j=1}^{b}\beta_j^2}{b-1}$ | $\sigma_e^2 + n\sigma_\gamma^2 + bn\sigma_\beta^2$ |
| Interaction | $SS_\gamma$ | $(a-1)(b-1)$ | $MS_\gamma$ | $\sigma_e^2 + n\sigma_\gamma^2$ | $\sigma_e^2 + n\sigma_\gamma^2$ |
| Error | $SS_e$ | $ab(n-1)$ | $MS_e$ | $\sigma_e^2$ | $\sigma_e^2$ |

not always the case. Sometimes it becomes necessary to use a test statistic that is a ratio of the sum of appropriate mean squares, both in the numerator and in the denominator. The test is then carried out using an approximate $F$ distribution. This procedure is known as the *Satterthwaite approximation*.

### *Analysis of Unbalanced Data*

The nice formulas and procedures available for mixed- and random-effects models with balanced data are not available in the case of unbalanced data. While some exact procedures can be derived in the case of a single factor experiment, that is, Model 1, such is not the case when we have a multifactor experiment. One option is to analyze the data with likelihood-based procedures. That is, one can estimate the parameters by maximizing the likelihood and then test the relevant hypotheses with likelihood ratio tests. The computations have to be carried out by available software.

As for estimating the random-effects variances, a point to note is that estimates can be obtained on the basis of either the likelihood or the *restricted likelihood*. Restricted likelihood is free of the fixed-effects parameters. The resulting estimates of the variances are referred to as *restricted maximum likelihood* (REML) *estimates*. REML estimates are preferred to maximum likelihood estimates because the REML estimates reduce (or eliminate) the bias in the estimates.

### Software for Data Analysis Based on Mixed- and Random-Effects Models

Many of the popular statistical software packages can be used for data analysis based on mixed- and random-effects models: R, S-Plus, SAS, SPSS (an IBM company, formerly called PASW® Statistics), Stata, and so forth. Their use is rather straightforward; in fact many excellent books are now available that illustrate the use of these software packages. A partial list of such books is provided in the Further Readings. These books provide the necessary software codes, along with worked-out examples.

*Thomas Mathew*

***See also*** Analysis of Variance (ANOVA); Experimental Design; Factorial Design; Fixed-Effects Models; Random-Effects Models; Simple Main Effects

**Further Readings**

Littell, R. C., Milliken, G. A., Stroup, W. W., & Wolfinger, R. D. (1996). *SAS system for mixed models*. Cary, NC: SAS Publishing.

Montgomery, D. C. (2009). *Design and analysis of experiments* (7th ed.). New York: Wiley.

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-Plus*. New York: Springer-Verlag.

Verbeke, G., & Molenberghs, G. (1997). *Linear mixed models in practice: A SAS-oriented approach* (*Lecture notes in statistics*, Vol. 126). New York: Springer-Verlag.

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. New York: Springer-Verlag.

West, B. T., Welch, K. B., & Galecki, A. T. (2007). *Linear mixed models: A practical guide using statistical software*. New York: Chapman & Hall/CRC.

## Mixed Methods Design

Mixed methods is a research orientation that possesses unique purposes and techniques. It

integrates techniques from quantitative and qualitative paradigms to tackle research questions that can be best addressed by mixing these two traditional approaches. As long as 40 years ago, scholars noted that quantitative and qualitative research were not antithetical and that every research process, through practical necessity, should include aspects of both quantitative and qualitative methodology. In order to achieve more useful and meaningful results in any study, it is essential to consider the actual needs and purposes of a research problem to determine the methods to be implemented. The literature on mixed methods design is vast, and contributions have been made by scholars from myriad disciplines in the social sciences. Therefore, this entry is grounded in the work of these scholars. This entry provides a historical overview of mixed methods as a paradigm for research, establishes differences between quantitative and qualitative designs, shows how qualitative and quantitative methods can be integrated to address different types of research questions, and illustrates some implications for using mixed methods. Though still new as an approach to research, mixed methods design is expected to soon dominate the social and behavioral sciences.

The objective of social science research is to understand the complexity of human behavior and experience. The task of the researcher, whose role is to describe and explain this complexity, is limited by his or her methodological repertoire. As tradition shows, different methods often are best applied to different kinds of research. Having the opportunity to apply various methods to a single research question can broaden the dimensions and scope of that research and perhaps lead to a more precise and holistic perspective of human behavior and experience. Research is not knowledge itself, but a process in which knowledge is constructed through step-by-step data gathering.

Data are gathered most typically through two distinct classical approaches—qualitative and quantitative. The use of both these approaches for a single study, although sometimes controversial, is becoming more widespread in social science. Methods are really "design" components that include the following: (a) the relationship between the researcher and research "subjects," (b) details of the experimental environment (place, time, etc.), (c) sampling and data collection methods, (d) data analysis strategies, and (e) knowledge dissemination. The design of a study thus leads to the choice of method strategy. The framework for a study, then, depends on the phenomenon being studied, with the participants and relevant theories informing the research design. Most study designs today need to include both quantitative and qualitative methods for gathering effective data and can thereby incorporate a more expansive set of assumptions and a broader worldview.

Mixing methods (or multiple-methods design) is generally acknowledged as being more pertinent to modern research than using a single approach. Quantitative and qualitative methods may rely more on single data collection methods. For example, whereas a quantitative study may rely on surveys for collecting data, a qualitative study may rely on observations or open-ended questions. However, it is also possible that each of these approaches may use multiple data collection methods. Mixed methods design "triangulates" these two types of methods. When these two methods are used within a single research study, different types of data are combined to answer the research question—a defining feature of mixed methods. This approach is already standard in most major designs. For example, in social sciences, interviews and participant observation form a large part of research and are often combined with other data (e.g., biological markers).

Even though the integration of these two research models is considered fairly novel (emerging significantly in the 1960s), the practice of integrating these two models has a long history. Researchers have often combined these methods, if perhaps only for particular portions of their investigations. Mixed methods research was more common in earlier periods when methods were less specialized and compartmentalized and when there was less orthodoxy in method selection. Researchers observed and cross-tabulated, recognizing that each methodology alone could be inadequate. Synthesis of these two classic approaches in data gathering and interpretation does not necessarily mean that they are wholly combined or that they are uniform. Often they need to be employed separately within a single research design so as not to corrupt either process.

Important factors to consider when one is using mixed methods can be summarized as follows. Mixed methods researchers agree that

there are some resonances between the two paradigms that encourage mutual use. The distinctions between these two methods cannot necessarily be reconciled. Indeed, this "tension" can produce more meaningful interactions and thus new results. Combination of qualitative and quantitative methods must be accomplished productively so that the integrity of each approach is not violated: Methodological congruence needs to be maintained so that data collection and analytical strategies are not jeopardized and can be consistent. The two seemingly antithetical research approaches can be productively combined in a pragmatic, interactive, and integrative design model. The two "classical" methods can complement each other and make a study more successful and resourceful by eliminating the possibility of distortion by strict adherence to a single formal theory.

## Qualitative and Quantitative Data

Qualitative and quantitative distinctions are grounded in two contrasting approaches to categorizing and explaining data. Different paradigms produce and use different types of data. Early studies distinguished the two methods according to the kind of data collected, whether textual or numerical. The classic qualitative approach includes study of real-life settings, focus on participants' context, inductive generation of theory, open-ended data collection, analytical strategies based on textual data, and use of narrative forms of analysis and presentation. Basically, the qualitative method refers to a research paradigm that addresses interpretation and socially constructed realities. The classic quantitative approach encompasses hypothesis formulation based on precedence, experiment, control groups and variables, comparative analysis, sampling, standardization of data collection, statistics, and the concept of causality. Quantitative design refers to a research paradigm that hypothesizes relationships between variables in an objective way.

Quantitative methods are related to deductivist approaches, positivism, data variance, and factual causation. Qualitative methods include inductive approaches, constructivism, and textual information. In general, quantitative design relies on comparisons of measurements and frequencies across categories and correlations between variables whereas the qualitative method concentrates on events within a context, relying on meaning and process. When the two are used together, data can be transformed. Essentially, "qualitized" data can represent data collected using quantitative methods that are converted into narratives that are analyzed qualitatively. "Quantitized" data represent data collected using qualitative methods that can be converted into numerical codes and analyzed statistically. Many research problems are not linear. Purpose drives the research questions. The course of the study, however, may change as it progresses, leading possibly to different questions and the need to alter method design. As in any rigorous research, mixed methods allows for the research question and purpose to lead the design.

## Historical Overview

In the *Handbook of qualitative research,* Norman K. Denzin and Yvonna S. Lincoln classified four historic periods in research history for the social sciences. Their classification shows an evolution from strict quantitative methodology, a gradual implementation and acceptance of qualitative methods, to a merging of the two: (1) traditional (quantitative), 1900 to 1950; (2) modernist, 1950 to 1970; (3) ascendance of constructivism, 1970 to 1990; and (4) pragmatism and the "compatibility thesis" (discussed later), 1990 to the present.

Quantitative methodology, and its paradigm, positivism, dominated methodological orientation during the first half of the 20th century. This "traditional" period, although primarily focused on quantitative methods, did include some mixed method approaches without directly acknowledging implementation of qualitative data: Studies often made extensive use of interviews and researcher observations, as demonstrated in the *Hawthorne effect.* In the natural sciences, such as biology, paleontology, and geology, goals and methods that typically would be considered qualitative (naturalistic settings, inductive approaches, narrative description, and focus on context and single cases) have been integrated with those that were regarded as quantitative (experimental mani-pulation, controls and variables, hypothesis testing, theory verification, and measurement and analysis of samples) for more than a century.

After World War II, positivism began to be discredited, which led to its "intellectual" successor, postpositivism. Postpositivism (still largely in the domain of the quantitative method) asserts that research data are influenced by the values of the researchers, the theories used by the researchers, and the researchers' individually constructed realities. During this period, some of the first explicit mixed method designs began to emerge. While there was no distinctive categorization of mixed methods, numerous studies began to employ components of its design, especially in the human sciences. Data obtained from participant observation (qualitative information) was often implemented, for example, to explain quantitative results from a field experiment.

The subsequent "modernist" period, or "Golden Age" (1950–1970), has been demarcated, then, by two trends: positivism's losing its stronghold and research methods that began to incorporate "multi methods." The discrediting of positivism resulted in methods that were more radical than those of postpositivism. From 1970 to 1985—defined by some scholars as the "qualitative revolution"—qualitative researchers became more vocal in their criticisms of pure quantitative approaches and proposed new methods associated with constructivism, which began to gain wider acceptance. In the years from 1970 to 1990, qualitative methods, along with mixed method syntheses, were becoming more eminent. In the 1970s, the combination of data sources and multiple methods was becoming more fashionable, and new paradigms, such as interpretivism and naturalism, were gaining precedence and validity.

In defense of a "paradigm of purity," a period known as the *paradigm wars* took place. Different philosophical camps held that quantitative and qualitative methods could not be combined; such a "blending" would corrupt accurate scientific research. Compatibility between quantitative and qualitative methods, according to these proponents of quantitative methods, was impossible due to the distinction of the paradigms. Researchers who combined these methods were doomed to fail because of the inherent differences in the underlying systems. Qualitative researchers defined such "purist" traditions as being based on "received" paradigms (paradigms preexisting a study that are automatically accepted as givens), and they argued against the prejudices and restrictions of positivism and postpositivism. They maintained that mixed methods were already being employed in numerous studies.

The period of pragmatism and compatibility (1990–the present) as defined by Denzin and Lincoln constitutes the establishment of mixed methods as a separate field. Mixed methodologists are not representative of either the traditional (quantitative) or "revolutionary" (qualitative) camps. In order to validate this new field, mixed methodologists had to show a link between epistemology and method and demonstrate that quantitative and qualitative methods were compatible. One of the main concerns in mixing methods was to determine whether it was also viable to mix paradigms—a concept that circumscribes an interface, in practice, between epistemology (historically learned assumptions) and methodology. A new paradigm, pragmatism, effectively combines these two approaches and allows researchers to implement them in a complementary way.

Pragmatism addresses the philosophical aspect of a paradigm by concentrating on what works. Paradigms, under pragmatism, do not represent the primary organizing principle for mixed methods practice. Believing that paradigms (socially constructed) are malleable assumptions that change through history, pragmatists make design decisions based on what is practical, contextually compatible, and consequential. Decisions about methodology are not based solely on congruence with established philosophical assumptions but are founded on a methodology's ability to further the particular research questions within a specified context. Because of the complexity of most contexts under research, pragmatists incorporate a dual focus between sense making and value making. Pragmatic research decisions, grounded in the actual context being studied, lead to a logical design of inquiry that has been termed *fitness for purpose*. Mixed methodologies are the result. Pragmatism demonstrates that singular paradigm beliefs are not intrinsically connected to specific methodologies; rather, methods and techniques are developed from multiple paradigms.

Researchers began to believe that the concept of a single best paradigm was a relic of the past and that multiple, diverse perspectives were critical to addressing the complexity of a pluralistic

society. They proposed what they defined as the *dialectical* stance: Opposing views (paradigms) are valid and provide for more realistic interaction. Multiple paradigms, then, are considered a foundation for mixed methods research. Researchers, therefore, need to determine which paradigms are best for a particular mixed methods design for a specific study.

Currently, researchers in social and behavioral studies generally comprise three groups: Quantitatively oriented researchers, primarily interested in numerical and statistical analyses; qualitatively oriented researchers, primarily interested in analysis of narrative data; and mixed methodologists, who are interested in working with both quantitative and qualitative data. The differences between the three groups (particularly between quantitatively and qualitatively oriented researchers) have often been characterized as the paradigm wars. These three movements continue to evolve simultaneously, and all three have been practiced concurrently. Mixed methodology is in its adolescent stage as scholars work to determine how to best integrate different methods.

## Integrated Design Models

A. Tashakkori and C. Teddlie have referred to three categories of multiple-method designs: multimethod research, mixed methods research, and mixed model research. The terms *multimethod* and *mixed method* are often confused, but they actually refer to different processes. In multimethod studies, research questions use both quantitative and qualitative procedures, but the process is applied principally to quantitative studies. This method is most often implemented in an interrelated series of projects whose research questions are theoretically driven. Multimethod research is essentially complete in itself and uses simultaneous and sequential designs.

Mixed methods studies, the primary concern of this entry, encompass both mixed methods and mixed model designs. This type of research implements qualitative and quantitative data collection and analysis techniques in parallel phases or sequentially. Mixed methods (combined methods) are distinguished from mixed model designs (combined quantitative and qualitative methods in all phases of the research). In mixed methods design, the "mixing" occurs in the type of questions asked and in the inferences that evolve. Mixed model research is implemented in all stages of the study (questions, methods, data collection, analysis, and inferences).

The predominant approach to mixing methods encompasses two basic types of design: *component* and *integrated*. In component designs, methods remain distinct and are used for discreet aspects of the research. Integrative design incorporates substantial integration of methods. Although typologies help researchers organize actual use of both methods, use of typologies as an organizing tool demonstrates a lingering linear concept that refers more to the duality of quantitative and qualitative methods than to the recognition and implementation of multiple paradigms. Design components (based on objectives, frameworks, questions, and validity strategies), when organized by typology, are perceived as separate entities rather than as interactive parts of a whole. This kind of typology illustrates a pluralism that "combines" methods without actually integrating them.

### *Triangulation and Validity*

Triangulation is a method that combines different theoretical perspectives within a single study. As applied to mixed methods, triangulation determines an unknown point from two or more known points, that is, collection of data from different sources, which improves validity of results. In *The Research Act,* Denzin argued that a hypothesis explored under various methods is more valid than one tested under only one method. Triangulation in methods, where differing processes are implemented, maximizes the validity of the research: Convergence of results from different measurements enhances validity and verification. It was also argued that using different methods, and possibly a faulty commonality of framework, could lead to increased error in results. Triangulation may not increase validity but does increase consistency in methodology: Though empirical results may be conflicting, they are not inherently damaging but render a more holistic picture.

Triangulation allows for the exploration of both theoretical and empirical observation (inductive and deductive), two distinct types of knowledge that can be implemented as a methodological

"map" and are logically connected. A researcher can structure a logical study, and the tools needed for organizing and analyzing data, only if the theoretical framework is established prior to empirical observations. Triangulation often leads to a situation in which different findings do not converge or complement each other. Divergence of results, however, may lead to additional valid explanations of the study. Divergence, in this case, can be reflective of a logical reconciliation of quantitative and qualitative methods. It can lead to a productive process in which initial concepts need to be modified and adapted to differing study results.

Recently, two new approaches for mixing methods have been introduced: an *interactive* approach, in which the design components are integrated and mutually influence each other, and a *conceptual* approach, using an analysis of the fundamental differences between quantitative and qualitative research. The interactive method, as employed in architecture, engineering, and art, is neither linear nor cyclic. It is a schematic method that addresses data in a mutually ongoing arrangement. This design model is a tool that focuses on analyzing the research question rather than providing a template for creating a study type. This more qualitative approach to mixed methods design emphasizes particularity, context, comprehensiveness, and the process by which a particular combination of qualitative and quantitative components develops in practice, in contrast to the categorization and comparison of data typical of the pure quantitative approach.

## Implications for Mixed Methods

As the body of research regarding the role of the environment and its impact on the individual has developed, the status and acceptance of mixed methods research in many of the applied disciplines is accelerating. This acceptance has been influenced by the historical development of these disciplines and an acknowledgment of a desire to move away from traditional paradigms of positivism and postpositivism. The key contributions of mixed methods have been to an understanding of individual factors that contribute to social outcomes, the study of social determinants of medical and social problems, the study of service utilization and delivery, and translational research into meaningful practice.

Mixed methods research may bridge postmodern critiques of scientific inquiry and the growing interest in qualitative research. Mixed methods research provides an opportunity to test research questions, hypotheses, and theory and to acknowledge the phenomena of human experience. Quantitative methods support the ability to generalize findings to the general population. However, quantitative approaches that are well regarded by researchers may not necessarily be comprehensible or useful to lay individuals. Qualitative approaches can help contextualize problems in narrative forms and thus can be more meaningful to lay individuals. Mixing these two methods offers the potential for researchers to understand, contextualize, and develop interventions.

Mixed methods have been used to examine and implement a wide range of research topics, including instrument design, validation of constructs, the relationship of constructs, and theory development or disconfirmation. Mixed methods are rooted, for one example, in the framework of feminist approaches whereby the study of participants' lives and personal interpretations of their lives has implications in research. In terms of data analysis, content analysis is a way for scientists to confirm hypotheses and to gather qualitative data from study participants through different methods (e.g., grounded theory, phenomenological, narrative). The application of triangulation methodology is extremely invaluable in mixed methods research.

While there are certainly advantages to employing mixed methods in research, their use also presents significant challenges. Perhaps the most significant issue to consider is the amount of time associated with the design and implementation of mixed methods. In addition to time restrictions, costs or barriers to obtaining funding to carry out mixed methods research are a consideration.

## Conclusion

Rather than choosing one paradigm or method over another, researchers often use multiple and mixed methods. Implementing these newer combinations of methods better supports the modern complexities of social behavior the changing

perceptions of reality and knowledge better serve the purposes of the framework of new studies in social science research. The classic quantitative and qualitative models alone cannot encompass the interplay between theoretical and empirical knowledge. Simply, combining methods makes common sense and serves the purposes of complex analyses. Methodological strategies are tools for inquiry and represent collections of strategies that corroborate a particular perspective. The strength of mixed methods is that research can evolve comprehensively and adapt to empirical changes, thus going beyond the traditional dualism of quantitative and qualitative methods, redefining and reflecting the nature of social reality.

Paradigms are social constructions, culturally and historically embedded as discourse practices, and contain their own set of assumptions. As social constructions, paradigms are changeable and dynamic. The complexity and pluralism of our contemporary world require rejecting investigative constraints of singular methods and implementing more diverse and integrative methods that can better address research questions and evolving social constructions. Knowledge and information change with time and mirror evolving social perceptions and needs. Newer paradigms and belief systems can help transcend and expand old dualisms and contribute to redefining the nature of social reality and knowledge.

Scholars generally agree that it is possible to use qualitative and quantitative methods to answer objective–value and subjective–constructivist questions, to include both inductive–exploratory and deductive–confirmatory questions in a single study, to mix different orientations, and to integrate qualitative and quantitative data in one or more stages of research, and that many research questions can only be answered with a mixed methods design. Traditional approaches meant aligning oneself to either quantitative or qualitative methods. Modern scholars believe that if research is to go forward, this dichotomy needs to be fully reconciled.

*Rogério M. Pinto*

***See also*** Critical Theory; Grounded Theory; Mixed Model Design; Qualitative Research; Quantitative Research

**Further Readings**

Creswell, J. W. (2003). *Research design: Qualitative, quantitative and mixed methods approach* (2nd ed.). Thousand Oaks, CA: Sage.

Denzin, N. K., & Lincoln, Y. S. (Eds.). (2000). *Handbook of qualitative research* (2nd ed.). Thousand Oaks, CA: Sage.

Tashakkori, A., & Teddlie, C. (Eds.). (2003). *Handbook of mixed methods in social and behavioral research*. Thousand Oaks, CA: Sage.

# Mixed Model Design

Mixed model designs are an extension of the general linear model, as in analysis of variance (ANOVA) designs. There is no common term for the mixed model design. Researchers sometimes refer to *split-plot designs, randomized complete block, nested, two-way mixed ANOVAs,* and certain repeated measures designs as mixed models. Also, mixed model designs may be restrictive or nonrestrictive. The restrictive model is used most often because it is more general, thus allowing for broader applications. A mixed model may be thought of as two models in one: a *fixed-effects model* and a *random-effects model*. Regardless of the name, statisticians generally agree that when interest is in both fixed and random effects, the design may be classified as a mixed model. Mixed model analyses are used to study research problems in a broad array of fields, ranging from education to agriculture, sociology, psychology, biology, manufacturing, and economics.

## Purpose of the Test

A mixed model analysis is appropriate if one is interested in a between-subjects effect (fixed effect) in addition to within-subjects effects (random effects), or in exploring alternative covariance structures on which to model data with between- and within-subjects effects. A variable may be fixed or random. Random effects allow the researcher to generalize beyond the sample.

Fixed and random effects are a major feature distinguishing the mixed model from the standard repeated measures design. The standard two-way repeated measures design examines repeated

measures on the same subjects. These are within-subjects designs for two factors with two or more levels. In the two-way mixed model design, two factors, one for within-subjects and one for between-subjects are always included in the model. Each factor has two or more levels. For example, in a study to determine the preferred time of day for undergraduate and graduate college students to exercise at a gym, time of day would be a within-subjects factor with three levels: 5:00 a.m., 1:00 p.m., and 9:00 p.m.; and student classification as undergraduate or graduate would be two levels of a between-subjects factor. The dependent variable for such a study could be a score on a workout preference scale. A design with three levels on a random factor and two levels on a fixed factor is written as a $2 \times 3$ mixed model design.

## Fixed and Random Effects

### *Fixed Effects*

Fixed effects, also known as between-subjects effects, are those in which each subject is a member of either one group or another, but not more than one group. All levels of the factor may be included, or only selected levels. In other words, subjects are measured on only one of the designated levels of the factor, such as undergraduate or graduate. Other examples of fixed effects are gender, membership in a control group or an experimental group, marital status, and religious affiliation.

### *Random Effects*

Random effects, also known as within-subjects effects, are those in which measures of each level of a factor are taken on each subject, and the effects may vary from one measure to another over the levels of the factor. Variability in the dependent variable can be attributed to differences in the random factor. In the previous example, all subjects would be measured across all levels of the time-of-day factor for exercising at a gym. In a study in which time is a random effect and gender is a fixed effect, the interaction of time and gender is also a random effect. Other examples of random effects are number of trials, in which each subject experiences each trial or each subject receives repeated doses of medication. Random effects are the measures from the repeated trials, measures after the time intervals of some activities, or repeated measures of some function such as blood pressure, strength level, endurance, or achievement. The mixed model design may be applied when the sample comprises large units, such as school districts, military bases, and universities, and the variability among the units, rather than the differences in means, is of interest. Examining random effects allows researchers to make inferences to a larger population.

## Assumptions

As with other inferential statistical procedures, the data for a mixed model analysis must meet certain statistical assumptions if trustworthy generalizations are to be made from the sample to the larger population. Assumptions apply to both the between- and within-subjects effects. The between-subjects assumptions are the same as those in a standard ANOVA: independence of scores; normality; and equal variances, known as *homogeneity of variance*. Assumptions for the within-subjects effects are independence of scores and normality of the distribution of scores in the larger population. The mixed model also assumes that there is a linear relationship between the dependent and independent variables. In addition, the complexity of the mixed model design requires the assumption of equality of variances of the difference scores for all pairs of scores at all levels of the within-subjects factor and equal covariances for the between-subjects factor. This assumption is known as the *sphericity assumption*. Sphericity is especially important to the mixed model analysis.

### *Sphericity Assumption*

The sphericity assumption may be thought of as the homogeneity-of-variance assumption for repeated measures. This assumption can be tested by conducting correlations between and among all levels of repeated measures factors and using Bartlett's test of sphericity. A significant probability level ($p$ value) means that the data are correlated and the sphericity assumption is violated. However, if the data are uncorrelated, then sphericity can be assumed. Multivariate ANOVA (MANOVA) procedures do not require that the sphericity

assumption be met; consequently, statisticians suggest using the MANOVA results rather than those in the standard repeated measures analysis.

Mixed model designs can accommodate complex covariance patterns for repeated measures with three or more levels on a factor, eliminating the need to make adjustments. Many different variance–covariance structures are available for fitting data to a mixed model design. Most statistical software can invoke the mixed model design and produce parameter estimates and tests of significance.

## Variance–Covariance Structures

Relationships between levels of the repeated measures are specified in a covariance structure. Several different covariance structures are available from which to choose. The variance–covariance structure of a data set is helpful to determine whether the data fit a specific model. A few of the most common covariance structures are introduced in the following paragraphs.

### *Diagonal, or Variance, Components Structure*

The diagonal, or variance, components structure is the simplest of the covariance patterns. This model, characterized by 1s on the diagonal and 0s on the off-diagonals, is also known as the *identity matrix*. The variance components pattern displays constant variance and no correlation between the elements. This is an unsatisfactory covariance structure for a mixed model design with repeated measures because the measures from the same subject are not independent.

### *Compound Symmetry*

Compound symmetry displays a covariance pattern among the multiple levels of a single factor in which all the off-diagonal elements are constant (equal) and all diagonal elements are constant (equal) regardless of the time lapse between measurements; thus such a pattern of equal variances and covariances indicates that the sphericity assumption is satisfied. The compound symmetry structure is a special case of the simple variance component model that assumes independent measures and homogeneous variance.

### *First-Order Autoregressive Structure*

The first-order autoregressive structure, or AR(1), model is useful when there are evenly spaced time intervals between the measurements. The covariance between any two levels is a function of the spacing between the measures. The closer the measurements are in time, the higher will be the correlations between adjacent measures. The AR(1) pattern displays a homogeneous structure, with equal variances and covariances that decrease exponentially as the time between measures increases. For example, this exponential function can be seen when an initial measure is taken and repeated at 1 year, 2 years, 3 years, and so forth. Smaller correlations will be observed for observations separated by 3 years than for those separated by 2 years. The model consists of a parameter for the variance of the observations and a parameter for the correlation between adjacent observations. The AR(1) is a special case of the Toeplitz, as discussed later in this entry.

### *Unstructured Covariance Structure*

The variances for each level of a repeated measures factor and the covariances in an unstructured variance–covariance matrix are all different. The unstructured covariance structure (UN), also known as a *general covariance structure,* is the most heterogeneous of the covariance structures. The UN possibly offers the best fit because every covariance entry can be unique. However, this heterogeneity introduces more complexity in the model and more parameters to be estimated. Unlike the first-order autoregressive structure that produces two parameters, the UN pattern requires parameters equal to $n(n+1)/2$, where $n$ is the number of repeated measures for a factor. For example, 10 parameters are estimated for four repeated measures: $4(4+1)/2 = 10$.

### *Toeplitz Covariance Structure*

In the Toeplitz covariance model, correlations at the same distance are the same as in the AR(1) model; however, there is no exponential effect. The number of parameters depends on the number of distances between measures. For example, the distance between the initial measure and the second measure would be one parameter; the distance

between the second measure and the third measure would be another parameter, and so forth. Like the AR(1) model, the Toeplitz is a suitable choice for evenly spaced measures.

### *First Order: Ante-Dependence*

The first-order ante-dependence model is a more general model than the Toeplitz or the AR(1) models. Covariances are dependent on the product of the variances at the two points of interest, and correlations are weighted by the variances of the two points of interest. For example, a correlation of .70 for points 1 and 2 and a correlation of .20 for points 2 and 3 would produce a correlation of .14 for points 1 and 3. This model requires $2n - 1$ parameters to be estimated, where $n$ is the number of repeated measures for a factor.

### *Evaluating Covariance Models*

The data should be examined prior to the analysis to verify whether the mixed model design or the standard repeated measures design is the appropriate procedure. Assuming that the mixed model procedure is appropriate for the data, the next step is to select the covariance structure that best models the data. The sphericity test alone is not an adequate criterion by which to select a model. A comparison of information criteria for several probable models with different covariance structures that uses the maximum likelihood and restricted maximum likelihood estimation methods is helpful in selecting the best model.

One procedure for evaluating a covariance structure involves creating a mixed model with an unstructured covariance matrix and examining graphs of the error covariance and correlation matrices. Using the residuals, the error covariances or correlations can be plotted separately for each start time, as in a trend analysis. For example, declining correlations or covariances with increasing time lapses between measures indicate that an AR(1) or ante-dependence structure is appropriate. For trend analysis, trends with the same mean have approximately the same variance. This pattern can also be observed on a graph with lines showing multiple trends. If the means or the lines on the graph are markedly different and the lines do not overlap, a covariance structure that accommodates variance heterogeneity is appropriate.

Another procedure for evaluating a covariance matrix involves creating several different probable models using both the maximum likelihood and the restricted maximum likelihood methods of parameter estimation. The objective is to select the covariance structure that gives the best fit of the data to the model. Information criterion measures, produced as part of the results of each mixed model procedure, indicate a relative goodness of fit of the data, thus providing guidance in model evaluation and selection. The information criteria measures for the same data set under different models (different covariance structures) and estimated with different methods can be compared; usually, the information criterion with the smallest value indicates a better fit of the data to the model. Several different criterion measures can be produced as part of the statistical analysis. It is not uncommon for the information criteria measures to be very close in value.

## Hypothesis Testing

The number of null hypotheses formulated for a mixed model design depends on the number of factors in the study. A null hypothesis should be generated for each factor and for every combination of factors. A mixed model analysis with one fixed effect and one random effect generates three null hypotheses. One null hypothesis would be stated for the fixed effects; another null hypothesis would be stated for the random effects; and a third hypothesis would be stated for the interaction of the fixed and random effects. If more than one fixed or random factor is included, multiple interactions may be of interest. The mixed model design allows researchers to select only the interactions in which they are interested.

The omnibus $F$ test is used to test each null hypothesis for mean differences across levels of the main effects and interaction effects. The sample means for each factor main effect are compared to ascertain whether the difference between the means can be attributed to the factor rather than to chance. Interaction effects are tested to ascertain whether a difference between the means of the fixed effects between subjects and the means of each level of the random effects within subjects is significantly different from zero. In other words,

the data are examined to ascertain the extent to which changes in one factor are observed across levels of the other factor.

## Interpretation of Results

Several tables of computer output are produced for a mixed model design. The tables allow researchers to check the fit of the data to the model selected and interpret results for the null hypotheses.

### *Model Dimension Table*

A model dimension table shows the fixed and random effects and the number of levels for each, type of covariance structure selected, and the number of parameters estimated. For example, AR(1) and compound symmetry covariance matrices estimate two parameters whereas the number of parameters varies for a UN based on the number of repeated measures for a factor.

### *Information Criteria Table*

Goodness-of-fit statistics are displayed in an information criteria table. Information criteria can be compared when different covariance structures and/or estimation methods are specified for the model. The tables resulting from different models can be used to compare one model with another. Information criteria are interpreted such that a smaller value means a better fit of the data to the model.

### *Fixed Effects, Random Effects, and Interaction Effects*

Parameter estimates for the fixed, random, and interaction effects are presented in separate tables. Results of the fixed effects allow the researcher to reject or retain the null hypothesis of no relationship between the fixed factors and the dependent variable. The level of significance ($p$ value) for each fixed effect will indicate the extent to which the fixed factor or factors have an effect different from zero on the dependent variable.

A table of estimates of covariance parameters indicates the extent to which random factors have an effect on the dependent variable. Random effects are reported as variance estimates. The level of significance allows the researcher to reject or retain the null hypothesis that the variance of the random effect is zero in the population. A nonsignificant random effect can be dropped from the model, and the analysis can be repeated with one or more other random effects.

Interaction effects between the fixed and random effects are also included as variance estimates. Interaction effects are interpreted on the basis of their levels of significance. For all effects, if the 95% confidence interval contains zero, the respective effects are nonsignificant. The residual parameter estimates the unexplained variance in the dependent variable after controlling for fixed effects, random effects, and interaction effects.

## Advantages

The advantages of the mixed model compensate for the complexity of the design. A major advantage is that the requirement of independence of individual observations does not need to be met as in the general linear model or regression procedures. The groups formed for higher-level analysis such as in nested designs and repeated measures are assumed to be independent; that is, they are assumed to have similar covariance structures. In the mixed model design, a wide variety of covariance structures may be specified, thus enabling the researcher to select the covariance structure that provides the model of best fit. Equal numbers of repeated observations for each subject are not required, making the mixed model design desirable for balanced and unbalanced designs. Measures for all subjects need not be taken at the same points in time. All existing data are incorporated into the analysis even though there may be missing data points for some cases. Finally, mixed model designs, unlike general linear models, can be applied to data at a lower level that are contained (nested) within a higher level, as in hierarchical linear models.

*Marie Kraska*

*See also* Hierarchical Linear Modeling; Latin Square Design; Sphericity; Split-Plot Factorial Design

**Further Readings**

Al-Marshadi, A. H. (2007). The new approach to guide the selection of the covariance structure in mixed model. *Research Journal of Medicine and Medical Sciences, 2*(2), 88–97.

Gurka, M. J. (2006). Selecting the best linear mixed model under REML. *American Statistician, 60*(1), 19–26.

Morrell, C. H., & Brant, L. J. (2000). Lines in random effects plots from the linear mixed-effects model. *American Statistician, 54*(1), 1–4.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear statistical models* (4th ed.). Boston: McGraw-Hill.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics, 57*(1), 120–125.

Schwarz, C. J. (1993). The mixed model ANOVA: The truth, the computer packages, the books. Part I: Balanced data. *American Statistician, 47*(1), 48–59.

VanLeeuwen, D. M. (1997). A note on the covariance structure in a linear model. *American Statistician, 51*(2), 140–144.

Voss, D. T. (1999). Resolving the mixed models controversy. *American Statistician, 53*(4), 352–356.

Wolfinger, R. (1993). Covariance structure selection in general mixed models. *Communications in Statistics—Simulation and Computation, 22,* 1079–1106.

# Mode

Together with the mean and the median, the mode is one of the main measurements of the central tendency of a sample or a population. The mode is particularly important in social research because it is the only measure of central tendency that is relevant for any data set. That being said, it rarely receives a great deal of attention in statistics courses. The purpose of this entry is to identify the role of the mode in relation to the median and the mean for summarizing various types of data.

## Definition

The mode is generally defined as the most frequent observation or element in the distribution. Unlike the mean and the median, there can be more than one mode. A sample or a population with one mode is unimodal. One with two modes is bimodal, one with three modes is trimodal, and so forth. In general, if there is more than one mode, one can say that a sample or a distribution is multimodal.

## History

The mode is an unusual statistic among those defined in the field of statistics and probability. It is a counting term, and the concept of counting elements in categories dates back prior to human civilization. Recognizing the maximum of a category, be it the maximum number of predators, the maximum number of food sources, and so forth, is evolutionarily advantageous.

The mathematician Karl Pearson is often cited as the first person to use the concept of the mode in a statistical context. Pearson, however, also used a number of other descriptors for the concept, including the "maximum of theory" and the "ordinate of maximum frequency."

## Calculation

In order to calculate the mode of a distribution, it is helpful first to group the data into like categories and to determine the frequency of each observation. For small samples, it is often easy to find the mode by looking at the results. For example, if one were to roll a die 12 times and get the following results,

$$\{1, 3, 2, 4, 6, 3, 4, 3, 5, 2, 5, 6\},$$

it is fairly easy to see that the mode is 3.

However, if one were to roll the die 40 times and list the results, the mode is less obvious:

$$\{6, 5, 5, 4, 4, 1, 6, 6, 3, 4, 4, 4, 2, 5,$$
$$5, 4, 4, 1, 2, 1, 4, 5, 5, 1, 3, 5, 2, 4,$$
$$2, 4, 2, 4, 4, 6, 5, 2, 1, 1, 4, 5\}.$$

In Table 1, the data are grouped by frequency, making it obvious that the mode is 4.

**Table 1** Data of Frequency of Die Rolled 40 Times

| *Number* | *Observations* |
|---|---|
| 1 | 6 |
| 2 | 6 |
| 3 | 2 |
| 4 | 13 |
| 5 | 9 |
| 6 | 4 |

**Table 2** Frequencies in Which Individual Data Are Grouped Into Ranges of Categories

| *Classes (income by categories)* | *Frequencies (number in each category)* |
|---|---|
| 0–$20k | 12 |
| $20k–$40k | 23 |
| $60k–$60k | 42 |
| $60k–$80k | 25 |
| $80k–$100k | 9 |
| $100k–$120k | 2 |
| Total | 113 |

Often, a statistician will be faced with a table of frequencies in which the individual data have been grouped into ranges of categories. Table 2 gives an example of such a table. We can see that the category of incomes between $40,000 and $60,000 has the largest number of members. We can call this the modal class. Yadolah Dodge has outlined a method for calculating a more precise estimate for the mode in such circumstances:

$$\text{mode} = L_1 + \left(\frac{d_1}{d_1 + d_2}\right) \times c,$$

where $L_1 =$ lower value of the modal category; $d_1 =$ difference between the number in the modal class and the class below; $d_2 =$ difference between the number in the modal class and the class above; and $c =$ length of the interval within the modal class. (This interval length should be common for all intervals.)

In this particular example, the mode would be

$$\text{mode} = \$40,000 + \left(\frac{19}{19 + 17}\right) \times \$20,000.$$

Therefore, the mode would be estimated to be $50,556.

## The Mode for Various Types of Data Classification

### *Discrete and Continuous Data*

In mathematical statistics, distributions of data are often discussed under the heading of discrete or continuous distributions. As the previous examples have shown, computing the mode is relatively straightforward with discrete data, although some organizing of the data may be necessary.

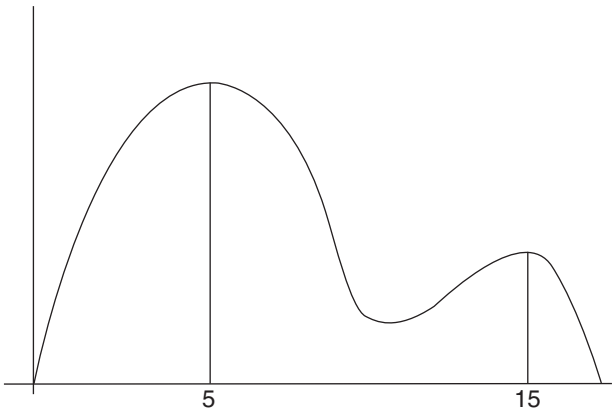



**Figure 1** Example of Modes for Continuous Distributions

For continuous data, the concept of mode is less obvious. No value occurs more than once in continuous probability distributions, and therefore, no single value can be defined as the mode with the discrete definition. Instead, for continuous distributions, the mode occurs at a local maximum in the data. For example, in Figure 1, the modes occur at 5 and 15.

As Figure 1 shows, for continuous distributions, a distribution can be defined as multimodal even when the local maximum at 5 is greater than the local maximum at 15.

### *The Stevens Classification System*

In social science statistical courses, data are often classified according to data scales outlined by Stanley Smith Stevens.

According to Stevens's classification system, data can be identified as nominal, ordinal, interval, or ratio. As Stevens points out, the mode is the only measure of central tendency applicable to all data scales.

Data that comply with *ratio scales* can be either continuous or discrete and are amenable to all forms of statistical analysis. In mathematical terms, data can be drawn from the real, integer, or natural number systems. Although the mode applies as a measure of central tendency for such data, it is not usually the most useful measurement of central tendency. An example of ratio-scale data

would be the total net worth of a randomly selected sample of individuals. So much variability is possible in the possible outcomes that unless the data are grouped into discrete categories (say increments of $5,000 or $10,000) the mode does not summarize the central tendency of the data well by itself.

*Interval data* are similar to ratio data in that it is possible to carry out detailed mathematical operations on them. As a result, it is possible to take the mean and the median as measures of central tendency. However, interval data lack a true 0 value.

For example, in measuring household size, it is conceivable that a household can possess very large numbers of members, but generally this is rare. Many households have fewer than 10 members; however, some modern extended families might run well into double digits. At the extreme, it is possible to observe medieval royal or aristocratic households with potentially hundreds of members. However, it is nonsensical to state that a household has zero members. It is also nonsensical to say that a specific household has 1.75 members. However, it is possible to say that the mean household size in a geographic region (a country, province, or city) is 2.75 or 2.2 or some other number. For interval data, the mode, the median, and the mean frequently provide valuable but different information about the central tendencies of the data. The mean may be heavily influenced by large low-prevalence values in the data; however, the median and the mode are much less influenced by them.

As an example of the role of the mode in summarizing interval data, if a researcher were interested in comparing the household size on different streets, A Street and B Street, he or she might visit both and record the following household sizes:

A Street: {3, 1, 6, 2, 1, 1, 2, 3, 2, 4, 2, 1, 4};
B Street: {2, 5, 3, 6, 7, 9, 3, 4, 1, 2, 1}.

Comparing the measures of central tendency (A Street: Mean = 2.5, Median = 2, Mode = 1 and 2; B Street: Mean = 3.8, Median = 3, Mode = 3) gives a clearer picture of the nature of the streets than any one measure of central tendency in isolation.

For data in *ordinal scales,* not only is there no absolute zero, but one can also rank the elements only in order of value. For example, a person could be asked to rank items on a scale of 1 to 5 in terms of his or her favorite. Likert-type scales are an example of this sort of data. A value of 5 is greater than a value of 4, but an increment from 4 to 5 does not necessarily represent the same increase in preference that an increase from 1 to 2 does. Because of these characteristics, reporting the median and the mode for this type of data makes sense, but the mean does not.

In the case of nominal-scale data, the mode is the only meaningful measure of central tendency. Nominal data tell the analyst nothing about the order of the data. In fact, data values do not even need to be labeled as numbers. One might be interested in which number a randomly selected group of hockey players at a hockey camp wear on their jersey back home in their regular hockey league. One could select from two groups:

Group 1: {1, 3, 4, 4, 4, 7, 7, 10, 11, 99};
Group 2: {1, 4, 8, 9, 11, 44, 99, 99, 99, 99}.

For these groups, taking the mean and the median are meaningless as measures of central tendency (the number 11 does not represent more value than 4). However, the mode of Group 1 is 4, and the mode of Group 2 is 99. With some background information about hockey, the analyst could hypothesize which populations the two groups are drawn from. Group 2 appears to be made up of a younger group of players whose childhood hero is Wayne Gretzky (number 99), and Group 1 is likely made up of an older group of fans of Bobby Orr (number 4).

Another interesting property of the mode is that the data do not actually need to be organized as numbers, nor do they need to be translated into numbers. For example, an analyst might be interested in the first names of CEOs of large corporations in the 1950s. Examining a particular newspaper article, the analyst might find the following names:

Names: {Ted, Gerald, John, Martin, John, Peter, Phil, Peter, Simon, Albert, John}

In this example, the mode would be *John,* with three listings. As this example shows, the mode is

particularly useful for textual analysis. Unlike the median and the mean, it is possible to take counts of the occurrence of words in documents, speech, or database files and to carry out an analysis from such a starting point.

Examination of the variation of nominal data is also possible by examining the frequencies of occurrences of entries. In the above example, it is possible to summarize the results by stating that John represents 3/11 (27.3%) of the entries, Peter represents 2/11 (18.2%) of the entries and that each other name represents 1/11 (9.1%) of the entries. Through a comparison of frequencies of occurrence, a better picture of the distribution of the entries emerges even if one does not have access to other measures of central tendency.

### A Tool for Measuring the Skew of a Distribution

Skew in a distribution is a complex topic; however, comparing the mode to the median and the mean can be useful as a simple method of determining whether data in a distribution are skewed. A simple example is as follows. In a workplace a company offers free college tuition for the children of its employees, and the administrator of the plan is interested in the amount of tuition that the program may have to pay. For simplicity, take three different levels of tuition. Private university tuition is set at \$30,000, out-of-state-student public university tuition is set at \$15,000, and in-state tuition is set at \$5,000.

In this example, 10 students qualify for tuition coverage for the current year. The distribution of tuition amounts for each student is as follows:

{5,000, 5,000, 5,000, 5,000, 5,000, 15,000, 15,000, 15,000, 30,000, 30,000}

Hence, the mode is \$5,000. The median is equal to $(15{,}000 + 5{,}000)/2 = \$10{,}000$. The mean = \$13,000. The tuition costs are skewed toward low tuition, and therefore the distribution has a negative skew.

A positively skewed distribution would occur if the tuition payments were as follows:

{30,000, 30,000, 30,000, 30,000, 30,000, 15,000, 15,000, 15,000, 5,000, 5,000}.

Now the mode is \$30,000, the median is $(30{,}000 + 15{,}000)/2 = \$22{,}500$, and the mean = \$20,500.

Finally, the following observations are possible:

{30,000, 30,000, 30,000, 15,000, 15,000, 15,000, 15,000, 15,000, 5,000, 5,000}.

In this case, the mode is \$15,000, the median is \$15,000, and the mean is \$17,500. The distribution has much less skew and is nearly symmetric.

The concept of skew can be dealt with using very complex methods in mathematical statistics; however, often simply reporting the mode together with the median and the mean is a useful method of forming first impressions about the nature of a distribution.

*Gregory P. Butler*

***See also*** Central Tendency, Measures of; Interval Scale; Likert Scaling; Mean; Median; Nominal Scale

**Further Readings**

Dodge, Y. (1993). *Statistique: Dictionnaire encyclopédique* [Statistics: Encyclopedic dictionary]. Paris: Dunod.

Magnello, M. E. (2009). Karl Pearson and the establishment of mathematical statistics. *International Statistical Review, 77*(1), 3–29.

Pearson, K. (1895). Contribution to the mathematical theory of evolution: Skew variation in homogenous material. *Philosophical Transactions of the Royal Society of London, 186*(1), 344–434.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*(2684), 677–680.

## Models

It used to be said that models were dispensable aids to formulating and understanding scientific theories, perhaps even props for poor thinkers. This negative view of the cognitive value of models in science contrasts with today's view that they are an essential part of the development of theories, and

more besides. Contemporary studies of scientific practice make it clear that models play genuine and indispensable cognitive roles in science, providing a basis for scientific reasoning. This entry describes types and functions of models commonly used in scientific research.

## Types of Models

Given that just about anything can be a model of something for someone, there is an enormous diversity of models in science. The many senses of the word *model* that stem from this bewildering variety Max Wartofsky has referred to as the "model muddle." It is not surprising, then, that the wide diversity of models in science has not been captured by some unitary account. However, philosophers such as Max Black, Peter Achinstein, and Rom Harré have provided useful typologies that impose some order on the variety of available models. Here, discussion is confined to four different types of model that are used in science: scale models, analogue models, mathematical models, and theoretical models.

### *Scale Models*

As their name suggests, scale models involve a change of scale. They are always models of something, and they typically reduce selected properties of the objects they represent. Thus, a model airplane stands as a miniaturized representation of a real airplane. However, scale models can stand as a magnified representation of an object, such as a small insect. Although scale models are constructed to provide a good resemblance to the object or property being modeled, they represent only selected relevant features of the object. Thus, a model airplane will almost always represent the fuselage and wings of the real airplane being modeled, but it will seldom represent the interior of the aircraft. Scale models are a class of iconic models because they literally depict the features of interest in the original. However, not all iconic models are scale models, as for example James Watson and Francis Crick's physical model of the helical structure of the DNA molecule. Scale models are usually built in order to present the properties of interest in the original object in an accessible and manipulable form. A scale model of an aircraft prototype, for example, may be built to test its basic aerodynamic features in a wind tunnel.

### *Analogue Models*

Analogue, or analogical, models express relevant relations of analogy between the model and the reality being represented. Analogue models are important in the development of scientific theories. The requirement for analogical modeling often stems from the need to learn about the nature of hidden entities postulated by a theory. Analogue models also serve to assess the plausibility of our new understanding of those entities.

Analogical models employ the pragmatic strategy of conceiving of unknown causal mechanisms in terms of what is already familiar and well understood. Well-known examples of models that have resulted from this strategy are the molecular model of gases, based on an analogy with billiard balls in a container; the model of natural selection, based on an analogy with artificial selection; and, the computational model of the mind, based on an analogy with the computer.

To understand the nature of analogical modeling, it is helpful to distinguish between a model, the source of the model, and the subject of the model. From the known nature and behavior of the source, one builds an analogue model of the unknown subject or causal mechanism. To take the biological example just noted, Charles Darwin fashioned his model of the subject of natural selection by reasoning analogically from the source of the known nature and behavior of the process of artificial selection. In this way, analogue models play an important creative role in theory development. However, this role requires the source from which the model is drawn to be different from the subject that is modeled. For example, the modern computer is a well-known source for the modeling of human cognition, although our cognitive apparatus is not generally thought to be a real computer. Models in which the source and the subject are different are sometimes called *paramorphs*. Models in which the source and the subject are the same are sometimes called *homeomorphs*. The paramorph can be an iconic, or pictorial, representation of real or imagined

things. It is iconic paramorphs that feature centrally in the creative process of theory development through analogical modeling.

In evaluating the aptness of an analogical model, the analogy between its source and subject must be assessed, and for this one needs to consider the structure of analogies. The structure of analogies in models comprises a positive analogy in which source and subject are alike, a negative analogy in which source and subject are unlike, and a neutral analogy where we have no reliable knowledge about matched attributes in the source and subject of the model. The negative analogy is irrelevant for purposes of analogical modeling. Because we are essentially ignorant of the nature of the hypothetical mechanism of the subject apart from our knowledge of the source of the model, we are unable to specify any negative analogy between the model and the mechanism being modeled. Thus, in considering the plausibility of an analogue model, one considers the balance of the positive and neutral analogies. This is where the relevance of the source for the model is spelled out.

An example of an analogue model is Rom Harré's rule-model of microsocial interaction, in which Erving Goffman's dramaturgical perspective provides the source model for understanding the underlying causal mechanisms involved in the production of ceremonial, argumentative, and other forms of social interaction.

### *Mathematical Models*

In science, particularly in the social and behavioral sciences, models are sometimes expressed in terms of mathematical equations. Mathematical models offer an abstract symbolic representation of the domain of interest. These models are often regarded as formalized theories in which the system modeled is projected on the abstract domain of sets and functions, which can be manipulated in terms of numerical reasoning, typically with the help of a computer. For example, factor analysis is a mathematical model of the relations between manifest and latent variables, in which each manifest variable is regarded as a linear function of a common set of latent variables and a latent variable that is unique to the manifest variable.

### *Theoretical Models*

Finally, the important class of models known as theoretical models abounds in science. Unlike scale models, theoretical models are constructed and described through use of the scientist's imagination; they are not constructed as physical objects. Further, unlike scale, analogical, and mathematical models, the properties of theoretical models are better known than the subject matter that is being modeled.

A theoretical model of an object, real or imagined, comprises a set of assumptions about that object. The Watson–Crick model of the DNA molecule and Markov models of human and animal learning are two examples of the innumerable theoretical models to be found in science. Theoretical models typically describe an object by ascribing to it an inner mechanism or structure. This mechanism or structure is frequently invoked in order to explain the behavior of the object. Theoretical models are acknowledged for their simplifying approximation to the object being modeled, and they can often be combined with other theoretical models to help provide a comprehensive understanding of the object.

## Data, Models, and Theories

### *Data Models*

In the 1960s, Patrick Suppes drew attention to the fact that science employs a hierarchy of models. He pointed out that theoretical models, which are high in the hierarchy, are not compared directly with empirical data, but with models of the data, which are lower in the hierarchy.

Data on their own are intractable. They are often rich, complex, and messy and, because of these characteristics, cannot be explained. Their intractability is overcome by reducing them into simpler and more manageable forms. In this way, data are reworked into models of data. Statistical methods play a prominent role in this regard, facilitating operations to do with assessing the quality of the data, the patterns they contain, and the generalizations to which they give rise. Because of their tractability, models of the data can be explained and used as evidence for or against theoretical models. For this reason, they are of considerable importance in science.

### *Models and Theories*

The relationship between models and theories is difficult to draw, particularly given that they can both be conceptualized in different ways. Some have suggested that theories are intended as true descriptions of the real world, whereas models need not be about the world, and therefore need not be true. Others have drawn the distinction by claiming that theories are more abstract and general than models. For example, evolutionary psychological theory can be taken as a prototype for the more specific models it engenders, such as those of differential parental investment and the evolution of brain size. Relatedly, Ronald Giere has argued that a scientific theory is best understood as comprising a family of models and a set of theoretical hypotheses that identify things in the world that apply to a model in the family.

Yet another characterization of models takes them to be largely independent of theories. In arguing that models are "autonomous agents" that mediate between theories and phenomena, Margaret Morrison contends that they are not fully derived from theory or data. Instead, they are technologies that allow one to connect abstract theories with empirical phenomena. Some have suggested that the idea of models as mediators does not apply to the behavioral and biological sciences because there is no appreciable gap between fundamental theory and phenomena in which models can mediate.

## The Functions of Models

### *Representation*

Models can variously be used for the purposes of systematization, explanation, prediction, control, calculation, derivation, and so on. In good part, models serve these purposes because they can often be taken as devices that represent parts of the world. In science, representation is arguably the main function of models. However, unlike scientific theories, models are generally not thought to be the sort of things that can be true or false. Instead, we may think of models as having a kind of similarity relationship with the object that is being modeled. With analogical models, for example, the similarity relationship is one of analogy. It can be argued that in science, models and theories are different representational devices. Consistent with this distinction between models and theories, William Wimsatt has argued that science often adopts a deliberate strategy of adopting false models as a means by which we can obtain truer theories. This is done by localizing errors in models in order eliminate other errors in theories.

### *Abstraction and Idealization*

It is often said that models provide a simplified depiction of the complex domains they often represent. The simplification is usually achieved through two processes: abstraction and idealization. Abstraction involves the deliberate elimination of those properties of the target that are not considered essential to the understanding of that target. This can be achieved in various ways; for example, one can ignore the properties, even though they continue to exist; one can eliminate them in controlled experiments; or one can set the values of unwanted variables to zero in simulations. By contrast, idealization involves transforming a property in a system into one that is related, but which possesses desirable features introduced by the modeler. Taking a spheroid object to be spherical, representing a curvilinear relation in linear form, and assuming that an agent is perfectly rational are all examples of idealization. Although the terms *abstraction* and *idealization* are sometime used interchangeably, they clearly refer to different processes. Each can take place without the other, and idealization can in fact take place without simplification.

*Brian D. Haig*

*See also* A Priori Monte Carlo Simulation; Exploratory Factor Analysis; General Linear Model; Hierarchical Linear Modeling; Latent Growth Modeling; Multilevel Modeling; Scientific Method; Structural Equation Modeling

**Further Readings**

Abrantes, P. (1999). Analogical reasoning and modeling in the sciences. *Foundations of Science, 4,* 237–270.

Black, M. (1962). *Models and metaphors: Studies in language and philosophy.* Ithaca, NY: Cornell University Press.

Giere, R. (1988). *Explaining science*. Chicago: University of Chicago Press.
Harré, R. (1976). The constructive role of models. In L. Collins (Ed.), *The use of models in the social sciences* (pp. 16–43). London: Tavistock.
MacCallum, R. C. (2003). Working with imperfect models. *Multivariate Behavioral Research, 38,* 113–139.
Morgan, M., & Morrison, M. (Eds.). (1999). *Models as mediators*. Cambridge, UK: Cambridge University Press.
Suppes, P. (1962). Models of data. In E. Nagel, P. Suppes, & A. Tarski (Eds.). *Logic, methodology, and philosophy of science: Proceedings of the 1960 International Congress* (pp. 252–261). Stanford, CA: Stanford University Press.
Wartofsky, M. (1979). *Models: Representation and the scientific understanding*. Dordrecht, the Netherlands: Reidel.
Wimsatt, W. C. (1987). False models as means to truer theories. In M. Nitecki & A. Hoffman (Eds.), *Neutral models in biology* (pp. 23–55). London: Oxford University Press.

## Monte Carlo Simulation

A Monte Carlo simulation is a methodological technique used to evaluate the empirical properties of some quantitative method by generating random data from a population with known properties, fitting a particular model to the generated data, collecting relevant information of interest, and replicating the entire procedure a large number of times (e.g., 10,000) in order to obtain properties of the fitted model under the specified condition(s). Monte Carlo simulations are generally used when analytic properties of the model under the specified conditions are not known or are unattainable. Such is often the case when no closed-form solutions exist, either theoretically or given the current state of knowledge, for the particular method under the set of conditions of interest. When analytic properties are known for a particular set of conditions, Monte Carlo simulation is unnecessary. Due to the computational tediousness of Monte Carlo methods because of the large number of calculations necessary, in practice they are essentially always implemented with one or more computers.

A Monte Carlo simulation study is a systematic investigation of the properties of some quantitative method under a variety of conditions in which a set of Monte Carlo simulations is performed. Thus, a Monte Carlo simulation study consists of the findings from applying a Monte Carlo simulation to a variety of conditions. The goal of a Monte Carlo simulation study is often to make general statements about the various properties of the quantitative method under a wide range of situations. So as to discern the properties of the quantitative method generally, and to search for inter-action effects in particular, a fully crossed factorial design is often used, and a Monte Carlo simulation is performed for each combination of the situations in the factorial design. After the data have been collected from the Monte Carlo simulation study, analysis of the data is necessary so that the properties of the quantitative procedure can be discerned. Because such a large number of replications (e.g., 10,000) are performed for each condition, the summary findings from the Monte Carlo simulations are often regarded as essentially population values, although confidence intervals for the estimates is desirable.

The general rationale of Monte Carlo simulations is to assess various properties of estimators and/or procedures that are not otherwise mathematically tractable. A special case of this is comparing the nominal and empirical values (e.g., Type I error rate, statistical power, standard error) of a quantitative method. Nominal values are those that are specified by the analyst (i.e., they represent the desired), whereas empirical values are those observed (i.e., they represent the actual) from the Monte Carlo simulation study. Ideally, the nominal and empirical values are equivalent, but this is not always the case. Verification that the nominal and empirical values are consistent can be the primary motivation for using a Monte Carlo simulation study.

As an example, under certain assumptions the standardized mean difference follows a known distribution, which in this case allows for exact analytic confidence intervals to be constructed for the population standardized mean difference. One of the assumptions on which the analytic procedure is based is that in the population, the scores within each of the two groups distribute normally. In order to evaluate the effectiveness of the (analytic)

approach to confidence interval formation when the normality assumption is not satisfied, Ken Kelley implemented a Monte Carlo simulation study and compared the nominal and empirical confidence interval coverage rates. Kelley also compared the analytic approach to confidence interval formation using two bootstrap approaches so as to determine whether the bootstrap performed better than the analytic approach under certain types of nonnormal data. Such comparisons require Monte Carlo simulation studies because no formula-based comparisons are available as the analytic procedure is based on the normality assumption, which was (purposely) not realized in the Monte Carlo simulation study.

As another example, under certain assumptions and an asymptotically large sample size, the sample root mean square error of approximation (RMSEA) follows a known distribution, which allows confidence intervals to be constructed for the population RMSEA. However, the effectiveness of the confidence interval procedure had not been well known for finite, and in particular small, sample sizes. Patrick Curran and colleagues have evaluated the effectiveness of the (analytic) confidence interval procedure for the population RMSEA by specifying a model with a known population RMSEA, generating data, forming a confidence interval for the population RMSEA, and replicating the procedure a large number of times. Of interest was the bias when estimating the population RMSEA from sample data and the proportion of confidence intervals that correctly bracketed the known population RMSEA, so as to determine whether the empirical confidence interval coverage was equal to the nominal confidence interval coverage (e.g., 90%).

A Monte Carlo simulation is a special case of a more general method termed the *Monte Carlo method*. The Monte Carlo method, in general, uses many sets of randomly generated data under some input specifications and applies a particular procedure or model to each set of the randomly generated data so that the output of interest from each fit of the procedure or model to the randomly generated data can be obtained and evaluated. Because of the large number of results of interest from the fitted procedure or model to the randomly generated data sets, the summary of the results describes the properties of how the procedure or model works in the specified input conditions.

A particular implementation of the Monte Carlo method is a method known as *Markov Chain Monte Carlo,* which is a method used to sample from various probability distributions based on a specified model in order to form sample means for approximating expectations. Markov Chain Monte Carlo techniques are most often used in the Bayesian approach to statistical inference, but they can also be used in the frequentist approach.

The term *Monte Carlo* was coined in the mid-1940s by Nicholas Metropolis while working at the Los Alamos National Laboratory with Stanislaw Ulam and John von Neumann, who proposed the general idea and formalized how determinate mathematical problems could be solved with random sampling from a specified model a large number of times, because of the games of chance commonly played in Monte Carlo, Monaco, with the idea of repeating a process a larger number of times and then examining the outcomes. The Monte Carlo method essentially replaced what was previously termed *statistical sampling*. Statistical sampling was used famously by William Sealy Gossett, who published under the name *Student,* before finalizing the statistical theory of the *t* distribution and was reported in his paper to show a comparison of empirical and nominal properties of the *t* distribution.

*Ken Kelley*

***See also*** A Priori Monte Carlo Simulation; Law of Large Numbers; Normality Assumption

**Further Readings**

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Currran, P. J., Bollen, K. A., Chen, F., Paxton, P., & Kirby, J. B. (2003). Finite sampling properties of the point estimates and confidence intervals of the RMSEA. *Sociological Methods Research, 32,* 208–252.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). Introducing Markov chain Monte Carlo. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter

(Eds.), *Markov chain Monte Carlo in practice* (pp. 1–20). New York: Chapman & Hall.
Kelley, K. (2005). The effects of nonnormal distributions on confidence intervals around the standardized mean difference: Bootstrap and parametric confidence intervals. *Educational & Psychological Measurement, 65,* 51–69.
Metropolis, N. (1987). The beginning of the Monte Carlo method. *Los Alamos Science,* 125–130.
Student. (1908). The probable error of the mean. *Biometrika, 6,* 1–25.

# Mortality

Mortality refers to death as a study endpoint or outcome. Broader aspects of the study of death and dying are embraced in the term *thanatology*. *Survival* is an antonym for *mortality*. Mortality may be an outcome variable in populations or samples, associated with treatments or risk factors. It may be a confounder of other outcomes due to resultant missing data or to biases induced when attrition due to death results in structural changes in a sample. Mortality is an event that establishes a metric for the end of the life span. Time to death is frequently used as an outcome and, less frequently, as a predictor variable. This entry discusses the use and analysis of mortality data in research studies.

## Population Mortality Rates

Nearly all governments maintain records of deaths. Thus many studies of mortality are based on populations rather than samples. The most common index of death in a specific group is its mortality rate. Interpretation of a mortality rate requires definition of the time, causes of death, and groups involved. Mortality rates are usually specified as the number of deaths in a year per 1,000 individuals, or in circumstances where mortality is rarer, per 100,000 individuals. A mortality rate may be cause specific, that is, refer to death due to a single condition, such as a disease or type of event or exposure. All-cause mortality refers to all deaths regardless of their cause. Mortality rates are often calculated for whole populations but can be expected to vary as a function of demographic variables, notably sex and age. Important subgroup mortality rates, as recognized by the World Health Organization, include the neonatal mortality rate, or deaths during the first 28 days of life per 1,000 live births; the infant mortality rate, or the probability of a child born in a specific year or period dying before reaching the age of 1 year; and the maternal mortality rate, or the number of maternal deaths due to childbearing per 100,000 live births. The adult mortality rate refers to death rate between 15 and 60 years of age. Age-specific mortality rates refer to the number of deaths in a year (per 100,000 individuals) for individuals of a certain age bracket. In comparing mortality rates between groups, age and other demographics must be borne in mind. Mortality rates may be standardized to adjust for differences in the age distributions of populations.

## Use in Research Studies

Mortality and survival are central outcomes in a variety of research settings.

### *Clinical Trials*

In clinical trials studying treatments for life-threatening illnesses, survival rate is often the primary outcome measure. Survival rate is evaluated as 1 minus the corresponding mortality rate. Randomized controlled trials are used to compare survival rates in patients receiving a new treatment to that in patients receiving a standard or placebo treatment. The latter is commonly known as the *control group*. Such trials should be designed to recruit sufficient numbers of patients and to follow them for long enough to observe deaths likely to occur due to the illness to ensure adequate statistical power to detect differences in rates.

### *Epidemiological Studies*

Epidemiological studies of mortality compare rates of death across different groups defined by demographic measures, by risk factors, by exposures, or by location.

#### Prospective Studies

Many studies make an initial assessment of a sample of interest and follow up with participants

at least once, but often at multiple, regular intervals. Such studies are in a strong position to make causal inferences by examining the association of previously measured variables with survival outcome. A variation on this design has measurement occasions defined by an event such as an episode of illness or a change in personal circumstances (for example, unemployment, marriage, parenthood, criminal conviction). In studies of this type, the death of participants may be an outcome of interest or a source of missing data.

#### Retrospective Studies

It is also possible to conduct retrospective studies using previously collected data to examine risk factors for mortality. For example, a previously conducted cross-sectional study may be augmented with the current vital status information of participants.

#### Case–Control Studies

Case–control studies involve recruiting a sample of participants who are classified on the basis of a certain condition (cases) and a separate sample of participants who do not meet the requirements of the condition (controls). Frequently cases are individuals diagnosed with an illness who are still living, but this design can be applied to those who have died from a specific cause. When the cases and controls are appropriately matched on confounding variables, the odds of risk factor status in cases versus that in controls can be calculated. This odds ratio can be shown to estimate relative risk of illness given exposure. This design is particularly useful for studying risk factors for uncommon conditions and associated mortality where it would be difficult to find a sufficiently large number of cases in an acceptable period.

## Assessment of Vital Status

Vital status is an indicator for the living status of an individual; that is, it indicates whether a person is alive or dead at a particular time. In studies that incorporate mortality as one of their study endpoints, information on vital status of participants must be collected and updated.

### *Source of Information*

In many countries, registries of deaths may be accessed by researchers and searched by name and date of birth. Newspaper death notices may also be a source of information. Ideally, multiple methods of tracking the vital status of participants should be used to ensure that deaths are not missed. It is critical that local legal and ethical requirements are adhered to when data or personal information is used in this way. In prospective studies, vital status may be collected when follow-up interviews are arranged and at any other points of contact.

### *Follow-Up of Subjects*

Studies of mortality frequently need to follow up with participants, often over an extended time, to accrue sufficient deaths for statistical power and rigor. The amount of time required will depend on the population being studied. Prospective studies, particularly clinical trials, are commonly designed to follow participants from the time they enter the study until either the time of death or the study's end. When definitive ascertainment of vital status cannot be achieved, uncertainty remains about whether the participant is dead, has withdrawn from the study or has simply lost contact.

In retrospective studies, it may be more difficult to track down participants, as their current information may be incomplete or may have changed over time. Participants may have moved or changed names, which makes tracking more difficult, especially when such information is not updated on a regular basis. Losing track of participants can lead to problems of response bias.

In prospective studies, researchers should collect sufficient tracking information at study commencement, such as comprehensive contact information for the participant and for their associates (e.g., relatives, friends, medical providers). Tracking information can be updated at each data collection juncture. The drawback of prospective studies, however, is that following participants over long periods can be very costly in terms of time and money. In the case of retrospective studies, tracking is performed only once based on existing information, causing minimal time and financial burden.

## Analysis of Mortality Data

### *Statistical Methods*

In mortality studies, the length of time from study entry to death, commonly known as *survival time,* is frequently specified as the primary outcome measure, with participants being followed over time on their vital status. However, due to time constraints and limited resources, studies are designed to follow participants for a predetermined period. Thus not all participants will be followed until death. This gives rise to *censored data.* For participants who are still alive at the end of a study, all that can be said is they have survived at least as long as the period of observation or follow-up and are not yet dead. Sophisticated statistical methods are required in the presence of censored data. Collectively known as *survival analysis,* these techniques may be applied for any time-to-event analyses, but they originated and are strongly associated with the modeling and prediction of time to death. A variety of techniques have been developed, including *Kaplan–Meier plots, Cox proportional hazards models,* and *frailty models.*

Survival time is usually modeled using Cox proportional hazards regression models, which take into account both survival time and whether the observation is censored. In Cox regression, survival time is considered a continuous outcome. Many statistical software packages also provide methods for modeling survival time that is measured discretely (for example, when only wave of measurement is available). An alternative method is to simultaneously model additional outcomes, such as dementia or institutionalization, with mortality in a latent random-effects model, called a frailty model. Latent growth models—a form of structural equation modeling—are another way of examining relationships between risk factors over time when modeling survival. Time to death can also be used as a predictor variable, in place of or in addition to chronological age or time in study. This measure may be a more useful indicator of biological aging than age (time since birth) in studies of late-life physical and cognitive function.

### *Mortality and Missing Data*

In longitudinal studies in which mortality is not the outcome of interest, participant mortality may be an important missing data problem. This applies particularly to studies of older or unwell samples. Survivor bias can occur when individuals with an attribute or risk factor of interest have a higher mortality rate than other members of the sample. This can result in a risk factor for early mortality appearing to be protective against the development of diseases of old age, such as Alzheimer's disease. In this case, investigators can evaluate the chance of developing Alzheimer's disease using a competing risks analysis incorporating death as a competing event.

Methods of handling missingness due to other factors may be applied equally when mortality is the cause of missing observations. There is no inherent objection to the use of multiple imputation to create values of deceased participants. Of more concern is whether the assumptions underlying the handling of missing observations are met. At best these techniques allow missingness to be associated with observed variables but not with the status of the missing observations themselves (missing at random). In many situations this assumption is not likely to be met, and the missingness mechanism is thus nonignorable and must be explicitly modeled. These procedures involve specifying mechanisms regarding the relationship between mortality and missing observations that cannot be evaluated from the data themselves. Often the results of such analyses are speculative and should be accompanied by sensitivity analyses.

*Philip J. Batterham, Andrew J. Mackinnon, and Kally Yuen*

***See also*** Clinical Trial; Cohort Design; Last Observation Carried Forward; Latent Growth Modeling; Longitudinal Design; Missing Data, Imputation of; Mixed- and Random-Effects Models; Survival Analysis

### Further Readings

Dufouil, C., Brayne, C., & Clayton, D. (2004). Analysis of longitudinal studies with death and drop-out: A case study. *Statistics in Medicine, 23,* 2215–2226.

Harel, O., Hofer, S. M., Hoffman, L., Pedersen, N. L., & Johansson, B. (2007). Population inference with mortality and attrition in longitudinal studies on aging: A two-stage multiple imputation method. *Experimental Aging Research, 33,* 187–203.

Hosmer, D. W., Lemeshow, S., & May, S. (2008). *Applied survival analysis: Regression modeling of time to event data*. New York: Wiley Interscience.

Ripatti, S., Gatz, M., Pedersen, N. L., & Palmgren, J. (2003). Three-state frailty model for age at onset of dementia and death in Swedish twins. *Genetic Epidemiology, 24,* 139–149.

Schlesselman, J. J. (1982). *Case-control studies: Design, conduct, analysis*. New York: Oxford University Press.

Whalley, L. J., & Deary, I. J. (2001). Longitudinal cohort study of childhood IQ and survival up to age 76. *British Medical Journal, 322,* 819–822.

World Health Organization. (2008). *WHO Statistical Information System: Indicator definitions and metadata, 2008*. Retrieved August 15, 2008, from http://www.who.int/whosis/indicators/compendium/2008/en

# Multilevel Modeling

Multilevel modeling (MLM) is a regression-based approach for handling nested and clustered data. *Nested data* (sometimes referred to as person–period data) occurs when research designs include multiple measurements for each individual, and this approach allows researchers to examine how participants differ, as well as how individuals vary across measurement periods. A good example of nested data is repeated measurements taken from people over time; in this situation, the repeated measurements are nested under each person. *Clustered data* involves a hierarchical structure such that individuals in the same group are hypothesized to be more similar to each other than to other groups. A good example of clustered data is the study of classrooms within different schools; in this situation, classrooms are embedded within the schools. Standard (ordinary least squares [OLS]) regression approaches assume that each observation in a data set is independent. Thus, it is immediately obvious that nested and hierarchically structured data violate this assumption of independence. MLM techniques arose to address this limitation of OLS regression. As discussed below, however, most of the common MLM techniques are extensions of OLS regression and are accessible to anyone with a basic working knowledge of multiple regression.

## History and Advantages

Statistical analyses conducted within an MLM framework date back to the late 19th century and the work of George Airy in astronomy, but the basic specifications used today were greatly advanced in the 20th century by Ronald Fisher and Churchill Eisenhart's introduction of fixed- and random-effects modeling. MLM permits the analysis of interdependent data without violating the assumptions of standard multiple regression. A critical statistic for determining the degree of interrelatedness in one's data is the intraclass correlation (ICC). The ICC is calculated as the ratio of between-group variance to between-subject variance, divided by total variance. The degree to which the ICC affects alpha levels is dependent on the size of a sample; small ICCs inflate alpha in large samples, whereas large ICCs will inflate alpha in small samples. A high ICC suggests that the assumption of independence is violated. When the ICC is high, using traditional methods such as multiple linear regression is problematic because ignoring the interdependence in the data will often yield biased results by artificially inflating the sample size in the analysis, which can lead to statistically significant findings that are not based on random sampling. In addition, it is important to account for the nested structure of the data—that is, nonindependence—to generate an accurate model of the variation in the data that is due to differences between groups and between subjects after accounting for within differences within groups and within subjects. Because variation within groups and within individuals usually accounts for most of the total variance, disregarding this information will bias these estimates.

In addition to its ability to handle nonindependent data, an advantage of MLM is that more traditional approaches for studying repeated measures, such as repeated measures analysis of variance (ANOVA), assume that data are completely balanced with, for example, the same number of students per classroom or equivalent measurements for each individual. Missing data or unbalanced designs cannot be accommodated with repeated measures ANOVA and are dropped from further analysis. MLM techniques were designed to use an iterative process of model estimation by which all data can be used in analysis; the two

most common approaches are maximum and restricted maximum likelihood estimation (both of which are discussed later in this entry).

## Important Distinctions

Multilevel models are also referred to as hierarchical linear models, mixed models, general linear mixed models, latent curve growth models, variance components analysis, random coefficients models, or nested or clustered models. These terms are appropriate and correct, depending on the field of study, but the multiple names can also lead to confusion and apprehension associated with using this form of statistical analysis. For instance, the hierarchical linear model 6.0, developed by Tony Bryk and Steve Raudenbush, is a statistical program that can handle both nested and clustered data sets, but MLM can be conducted in other popular statistical programs, including SAS, SPSS (an IBM company, formerly called PASW® Statistics), and R, as well as a host of software for analyzing structural equation models (e.g., LISREL, MPlus). MLM can be used for nonlinear models as well, such as those associated with trajectories of change and growth, which is why the terms *hierarchical linear modeling* and *general linear mixed models* can be misleading. In addition, latent curve growth analysis, a structural equation modeling technique used to fit different curves associated with individual trajectories of change, is statistically identical to regression-based MLM. The different approaches recognize that data is not always nested within a hierarchical structure, and also that data may exhibit nonlinear trajectories such as quadratic or discontinuous change. MLM also is referred to as mixed models analysis by researchers interested in the differences between subjects and groups and within subjects and groups that account for variance in their data. Finally, variance components analysis and random coefficients models refers to variance that is assumed to be random across groups or individuals as opposed to fixed, as is assumed in single-level regression; MLM is referred to by this terminology as well.

## Person–Period and Clustered Data

As mentioned, MLM can be used flexibly with multiple types of data structures. Two of the most common structures are person–period data and clustered data. Person–period data examine both between- and within-individual variation, with the latter examining how an individual varies across a measurement period. Most studies having this design include longitudinal data that examines individual growth. An example might be a daily diary study in which each diary entry or measurement (Level 1) is nested within an individual (Level 2). Ostensibly, a researcher might be interested in examining change within an individual's daily measurements across time or might be investigating how daily ratings differ as a function of between-individual factors such as personality, intelligence, age, and so forth. Thus, at the within-person level (Level 1), there may be predictors associated with an individual's rating at any given occasion, but between-person (Level 2) variables may also exist that moderate the strength of that association. As opposed to assuming that individuals' responses are independent, the assumption of MLM is that these responses are inherently related and more similar within an individual than they are across individuals.

A similar logic applies for clustered data. A common example of hierarchically structured data in the education literature assumes that students in the same classroom will be more similar to each other than to students in another class. This might result from being exposed to the same teacher, materials, class activities, teaching approach, and so forth. Thus, students' individual responses (Level 1) are considered nested within classrooms (Level 2). A researcher might be interested in examining individual students' performance on an arithmetic test to ascertain if variability is due to differences among students (Level 1) or between classrooms (Level 2). Similar to its use with person–period data, MLM in this example can investigate within-classroom variability at the lowest level and between-classroom variability at the highest level of the hierarchy.

The main distinction between these data structures rests in the information that is gleaned for data analysis. For person–period data, one can make inferences regarding variability within person responses or trajectories of change over a time, which may help answer questions related to the study of change. For clustered data, one can study differences among and within groups, which may

help answer questions regarding program evaluation. Perhaps not surprisingly, these two structures also can be combined, such as when multiple arithmetic exams are given over time (Level 1) and nested within each student (Level 2) who remains assigned to a classroom (Level 3). A thorough discussion of this three-level example is beyond the scope of the present entry, but the topic is raised as an example of the flexibility and sophistication of MLM.

## The Multilevel Model

This discussion follows the formal notation introduced by Bryk and Raudenbush, Judith Singer, John Willet, and others. A standard two-level equation for the lower and higher levels of a hierarchy that includes a predictor at Level 1 is presented first, followed by a combined equation showing the collapsed single-level model. This last step is important because, depending on which software is chosen, the two-level model (e.g., HLM 6.0) or the collapsed model (e.g., SAS Proc-Mixed) may require an explicit equation. It is important to note, also, that these equations are the same for any two-level person–period or clustered data set.

The two-level model is presented below in its simplest form:

$$\text{Level 1}: Y_{ij} = \beta 0_i + \beta 1_i + \varepsilon_{ij}, \tag{1}$$

where $i$ refers to individual and $j$ refers to time, $\beta 0_i$ is the intercept for this linear model, and $\beta 1_i$ is the slope for the trajectory of change. Notice that the Level 1 equation looks almost identical to the equation used for a simple linear regression. The main differences are the error term $\varepsilon_{ij}$ and the introduction of subscripts $i$ and $j$. The error term signifies random measurement error associated with data that, contrary to the slope, deviate from linearity. For the earlier daily diary (individual–period) example, the Level 1 equation details that individual $i$'s rating at time $j$ is dependent on his or her first rating ($\beta 0_i$) and the slope of linear change ($\beta 1_i$) between time (or occasion) 1 and time $j$ (note that $\beta 1_i$ does not always represent time, but more generally represents a 1-unit change from baseline in the time-varying Level 1 predictor; the equations are identical, however). The Level 2 equation estimates the individual growth parameters of the intercept and slope for each individual by the following equations:

$$\begin{aligned}\text{Level 2}: \beta 0_i &= \gamma 00 + \zeta 0_i \\ \beta 1_i &= \gamma 10 + \zeta 1_i(2)\end{aligned}$$

where $\zeta 0_i$ and $\zeta 1_i$ indicate that the Level 2 outcomes ($\beta 0_i$ and $\beta 1_i$, the intercept and the slope from the Level 1 model) each have a residual term, while $\gamma 00$ represents the grand mean and $\gamma 10$ indicates the grand slope for the sample. This means that the intercept and slope are expected to vary across individuals and will deviate from the average intercept and slope of the entire sample.

By substituting the Level 2 equations into the Level 1 equation, one can derive the collapsed model:

$$\begin{aligned}Y_{ij} &= [(\gamma 00 + \zeta 0_i) + (\gamma 10 + \zeta 1_i)] + \varepsilon_{ij} \\ Y_{ij} &= (\gamma 00 + \gamma 10) + (\zeta 1_i + \zeta 0_i + \varepsilon_{ij}).\end{aligned} \tag{3}$$

## Types of Questions Answered Using Multilevel Models

This section details the most useful and common approaches for examining multilevel data and is organized in a stepwise fashion, with each consequent model adding more information and complexity.

### *Unconditional Means Model*

This approach is analogous to a one-way ANOVA examining the random effect, or variance in means, across individuals in a person–period data set, or across groups with clustered data. This model is run without any predictors at Level 1, which is equivalent to the model included in Equation 1 without any Level 1 predictors (i.e., the unconditional means model with only the intercept and error term). It is by running this model that one can determine the ICC and assess whether a multilevel analysis is indeed warranted. Thus, the unconditional means model provides an estimate of how much variance exists between groups and between subjects, as well as within groups and within subjects in the sample.

### *Means-as-Outcomes Model*

This model attempts to explain the variation that occurs in individual or group means as a function of a Level 2 variable. For example, perhaps it is believed that extroversion explains the mean differences in happiness people report across days, or that the number of pop quizzes administered will predict the mean arithmetic scores for a class. It is hypothesized that with inclusion of a Level 2 predictor, the individual or group mean will be altered compared with that of the unconditional means model. In order to ascertain whether this association holds true, one compares the difference in the between-subject and between-group variance in the unconditional means model with the difference in the between-subject and between-group variance in the means-as-outcomes models. One may also examine this model if significant between-groups error variance is exhibited in the unconditional means model because the model should be respecified to account for this Level 2 variation.

### *Random Coefficient Model*

As opposed to the means-as-outcomes model, the random coefficient model attempts to partition variability within groups and within subjects as a function of a Level 1 predictor. For the previous examples, this approach will yield the average intercept of happiness and the average slope of extroversion and happiness of all individuals for person–period data or the average intercept for arithmetic scores and the average slope of pop quizzes and arithmetic scores of all classrooms. In addition, this approach tests for significant differences between each individual and each classroom. Average intercept and slope estimates are yielded in the output as fixed effects, while a significant $p$ value for the variance of the slope estimate indicates that the slope varies across individuals and classrooms.

### *Intercepts and Slopes as Outcomes Model*

The intercepts and slopes as outcomes model includes both Level 1 and Level 2 predictors. This step is completed only if significant variability is accounted for by Level 2 and Level 1 predictors. As stated earlier, model testing occurs in a stepwise fashion, with this model serving as the final step.

## Error Variance and Covariance

The standard MLM for assessing change can be analyzed with simple multiple regression techniques, but it is important to recognize that a multilevel approach should use a specialized error term (i.e., a term that specifies why and in what way data vary). As presented in the collapsed model (Equation 3) above, the error term in this simplified regression equation has three components: the error associated with each Level 1 equation and two terms associated with the error around the intercept and the slope, respectively. Singer was among the first to point out that these associated error terms render OLS regression analysis unfit for nested data because it assumes that the exhibited variability should approximate a multivariate normal distribution. If this were true, the mean of the residuals would be zero, the residuals would be independent of one another (they would not covary), and the population variance of the residuals would be equivalent for each measurement. Nested data sets do not adhere to these assumptions; whereas it is expected that there will be independence among individuals and/or groups, it is also assumed that the variance within person and/or group will covary and each measurement and/or person will be correlated to the others. Specifically, for longitudinal, or person–period, data, special consideration must be given to the time dependence of the measurements for each person. There are multiple approaches that one can take when specifying how the error term should be structured in an MLM. The following is a brief and basic description of these various approaches.

### *Partitioning Variance Between Groups*

#### Random Intercept

This approach is also called the variance components, or between-groups compound symmetric structure. It is commonly used with longitudinal data in which subjects are allowed to vary and each subject has a mean intercept or score that deviates from the population (or all-subjects) mean. In this instance, the residual intercepts are thought to be normally distributed with a mean of zero.

#### Random Intercept and Slope

This approach is ideal for data in which subjects are measured on different time schedules, and it allows each subject to deviate from the population in terms of both intercept and slope. Thus, each individual can have a different growth trajectory even if the hypothesis is that the population will approximate similar shapes in its growth trajectory. Using this method, one can specify the residuals for the intercept and slopes to be zero or nonzero and alter the variance structures by group such that they are equal or unbalanced.

### *Partitioning Variance Within Groups*

#### Unstructured

An alternative approach is to estimate the within-group random effects. This approach assumes that each subject and/or group is independent with equivalent variance components. In this approach, the variance can differ at any time, and covariance can exist between all the variance components. This is the default within-groups variance structure in most statistical software packages and should serve as a starting point in model testing unless theory or experimental design favor another approach.

#### Within-Group Compound Symmetric

This approach constrains the variance and covariance to a single value. Doing so assumes that the variance is the same regardless of the time the individual was measured or the subject within the group and that the correlation between measurements will be equivalent. A subspecification of this error structure is the heterogeneous compound symmetric that dictates the variance is a single value, but the covariance between measurements can differ.

#### Autoregressive

Perhaps the most useful error structure for longitudinal data, this approach dictates that variance is the same at all times but that covariance decreases as measurement occasions are further apart. From a theoretical standpoint, this may not be the case in clustered data sets, but one can see how a person's ratings may be more similar on Days 1 and 2 compared with Days 1 and 15.

### *Model Estimation Methods*

As stated above, MLM differs from repeated measures ANOVA in that it uses all available data. To accomplish this, most statistical packages use a form of maximum likelihood (ML) estimation. ML estimations are favored because it is assumed that they converge on population estimates, that the sampling distribution is equivalent to the known variance, and that the standard error derived from the use of this method is smaller than from other approaches. These advantages apply only with large samples because ML estimations are biased toward large samples and their variance estimation may become unreliable with a smaller data set. There are two types of ML techniques, full ML and restricted ML. Full ML assumes that the dependent variable is normally distributed, and the mean is based on the regression coefficients and the variance components. Restricted ML uses the least squares residuals that remain after the influence of the fixed effects are removed and only the variance components remain. With ML algorithms, a statistic of fit is usually compared across models to reveal which model best accounts for variance in the dependent variable (e.g., the deviance statistic). Multiple authors suggest that when using restricted ML, one should make sure that models include the same fixed effects and that only the random effects vary, because one wants to make sure the fixed effects are accounted for equivalently across models.

A second class of estimation methods are extensions of OLS estimation. Generalized least squares (GLS) estimation allows the residuals to be autocorrelated and have more dispersion of the variances, but it requires that the actual amount of autocorrelation and dispersion be known in the population in order for one to accurately estimate the true error in the covariance matrix. In order to account for this, GLS uses the estimated error covariance matrix as the true error covariance matrix, and then it estimates the fixed effects and associated standard errors. Another approach, iterative GLS, is merely an extension of GLS and uses iterations that repeatedly estimate and refit the model until either the model is ideally converged or the

maximum number of iterations has occurred. This method also works only with large and relatively balanced data.

### Other Applications

The applications of MLM techniques are virtually limitless. Dyadic analytic methods are being used in MLM software, with individuals considered nested within dyads. Mediation and moderation analysis are possible both within levels of the hierarchy and across levels. Moderated mediation and mediated moderation principles, relatively new to the literature, also can be applied within a MLM framework.

### Resources

MLM workshops are offered by many private companies and university-based educational programs. Articles on MLM are easily located on academic databases. Another resource is the University of California–Los Angeles Stat Computing Portal, which has links to pages and articles of interest directly related to different aspects of MLM. Finally, those interested in exploring MLM will find it easily accessible as some programs are designed specifically for MLM analyses (e.g., HLM 6.0), but many of the more commonly used statistical packages (e.g., SAS and SPSS) have the same capabilities.

*Lauren A. Lee and David A. Sbarra*

***See also*** Analysis of Variance (ANOVA); Growth Curve; Hierarchical Linear Modeling; Intraclass Correlation; Latent Growth Modeling; Longitudinal Design; Mixed- and Random-Effects Models; Mixed Model Design; Nested Factor Design; Regression Artifacts; Repeated Measures Design; Structural Equation Modeling; Time-Series Study; Variance

#### Further Readings

Heck, R. H., & Thomas, S. L. (2000). *An introduction to multilevel modeling techniques.* Mahwah, NJ: Lawrence Erlbaum.

Kreft, I., & de Leeuw, J. (1998). *Introducing multilevel modeling.* Thousand Oaks, CA: Sage.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear model: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational & Behavioral Statistics, 23,* 323–355.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* New York: Oxford University Press.

Wallace, D., & Green, S. B. (2002). Analysis of repeated measures designs with linear mixed models. In D. S. Moskowitz & S. L. Hershberger (Eds.), *Modeling intraindividual variability with repeated measures data: Methods and applications* (pp. 103–134). Mahwah, NJ: Lawrence Erlbaum.

#### Websites

Scientific Software International. Hierarchical Linear and Nonlinear Modeling (HLM): http://www.ssicentral.com/hlm

UCLA Stat Computing Portal: http://statcomp.ats.ucla.edu

## Multiple Comparison Tests

Many research projects involve testing multiple research hypotheses. These research hypotheses could be evaluated using comparisons of means, bivariate correlations, regressions, and so forth, and in fact most studies consist of a mixture of different types of test statistics. An important consideration when conducting multiple tests of significance is how to deal with the increased likelihood (relative to conducting a single test of significance) of falsely declaring one (or more) hypotheses statistically significant, titled the *multiple comparisons problem.* This multiple comparisons problem is especially relevant to the topic of research design because the issues associated with the multiple comparisons problem relate directly to designing studies (i.e., number and nature of variables to include) and deriving a data analysis strategy for the study. This entry introduces the multiple comparisons problem and discusses some of the strategies that have been proposed for dealing with it.

### The Multiple Comparisons Problem

To help clarify the multiple comparisons problem, imagine a soldier who needed to cross fields containing land mines in order to obtain supplies. It is

clear that the more fields the individual crosses, the greater the probability that he or she will activate a land mine; likewise, researchers conducting many tests of significance have an increased chance of erroneously finding tests significant. It is important to note that although the issue of multiple hypothesis tests has been labeled the multiple comparisons problem, most likely because a lot of the research on multiple comparisons has come within the framework of mean comparisons, it applies to any situation in which multiple tests of significance are being performed.

Imagine that a researcher is interested in determining whether overall course ratings differ for lecture, seminar, or computer-mediated instruction formats. In this type of experiment, researchers are often interested in whether significant differences exist between any pair of formats, for example, do the ratings of students in lecture-format classes differ from the ratings of students in seminar-format classes. The multiple comparisons problem in this situation is that in order to compare each format in a pairwise manner, three tests of significance need to be conducted (i.e., comparing the means of lecture vs. seminar, lecture vs. computer-mediated, and seminar vs. computer-mediated instruction formats). There are numerous ways of addressing the multiple comparisons problem and dealing with the increased likelihood of falsely declaring tests significant.

### *Common Multiple Testing Situations*

There are many different settings in which researchers conduct null hypothesis testing, and the following are just a few of the more common settings in which multiplicity issues arise: (a) conducting pairwise and/or complex contrasts in a linear model with categorical variables; (b) conducting multiple main effect and interaction tests in a factorial analysis of variance (ANOVA) or multiple regression setting; (c) analyzing multiple simple effect, interaction contrast, or simple slope tests when analyzing interactions in linear models; (d) analyzing multiple univariate ANOVAs after a significant multivariate ANOVA (MANOVA); (e) analyzing multiple correlation coefficients; (f) assessing the significance of multiple factor loadings or factor correlations in factor analysis; (g) analyzing multiple dependent variables separately in linear models; (h) evaluating multiple parameters simultaneously in a structural equation model; and (i) analyzing multiple brain voxels for stimulation in functional magnetic resonance imaging research. Further, as stated previously, most studies involve a mixture of many different types of test statistics.

An important factor in understanding the multiple comparisons problem is understanding the different ways in which a researcher can "group" his or her tests of significance. For example, suppose, in the study looking at whether student ratings differ across instruction formats, that there was also another independent variable, the sex of the instructor. There would now be two "main effect" variables (instruction format and sex of the instructor) and potentially an interaction between instruction format and sex of the instructor. The researcher might want to group the hypotheses tested under each of the main effect (e.g., pairwise comparisons) and interaction (e.g., simple effect tests) hypotheses into separate "families" (groups of related hypotheses) that are considered simultaneously in the decision process. Therefore, control of the Type I error (error of rejecting a true null hypothesis) rate might be imposed separately for each family, or in other words, the Type I error rate for each of the main effect and interaction families is maintained at $\alpha$. On the other hand, the researcher may prefer to treat the entire set of tests for all main effects and interactions as one family, depending on the nature of the analyses and the way in which inferences regarding the results will be made. The point is that when researchers conduct multiple tests of significance, they must make important decisions about how these tests are related, and these decisions will directly affect the power and Type I error rates for both the individual tests and for the group of tests conducted in the study.

## Type I Error Control

Researchers testing multiple hypotheses, each with a specified Type I error probability ($\alpha$), risk an increase in the overall probability of committing a Type I error as the number of tests increases. In some cases, it is very important to control for Type I errors. For example, if the goal of the researcher comparing the three classroom instruction formats

described earlier was to identify the most effective instruction method that would then be adopted in schools, it would be important to ensure that a method was not selected as superior by chance (i.e., a Type I error). On the other hand, if the goal of the research was simply to identify the best classroom formats for future research, the risks associated with Type I errors would be reduced, whereas the risk of not identifying a possibly superior method (i.e., a Type II error) would be increased. When many hypotheses are being tested, researchers must specify not only the level of significance, but also the unit of analysis over which Type I error control will be applied. For example, the researcher comparing the lecture, seminar, and computer-mediated instruction formats must determine how Type I error control will be imposed over the three pairwise tests. If the probability of committing a Type I error is set at $\alpha$ for each comparison, then the probability that at least one Type I error is committed over all three pairwise comparisons can be much higher than $\alpha$. On the other hand, if the probability of committing a Type I error is set at $\alpha$ for all tests conducted, then the probability of committing a Type I error for each of the comparisons can be much lower than $\alpha$. The conclusions of an experiment can be greatly affected by the unit of analysis over which Type I error control is imposed.

## *Units of Analysis*

Several different units of analysis (i.e., error rates) have been proposed in the multiple comparison literature. The majority of the discussion has focused on the per-test and familywise error rates, although other error rates, such as the *false discovery rate,* have recently been proposed.

### *Per-Test Error Rate*

Controlling the per-test error rate ($\alpha_{PT}$) involves simply setting the $\alpha$ level for each test ($\alpha_T$) equal to the global $\alpha$ level. Recommendations for controlling $\alpha_{PT}$ center on a few simple but convincing arguments. First, it can be argued that the natural unit of analysis is the test. In other words, each test should be considered independent of how many other tests are being conducted as part of that specific analysis or that particular study or, for that matter, how many tests the researcher might conduct over his or her career. Second, real differences between treatment groups are more likely with a greater number of treatment groups. Therefore, emphasis in experiments should be not on controlling for unlikely Type I errors but on obtaining the most power for detecting even small differences between treatments. The third argument is the issue of (in)consistency. With more conservative error rates, different conclusions can be found regarding the same hypothesis, even if the test statistics are identical, because the per-test $\alpha$ level depends on the number of comparisons being made. Last, one of the primary advantages of $\alpha_{PT}$ control is convenience. Each of the tests is evaluated with any appropriate test statistic and compared to an $\alpha$-level critical value.

The primary disadvantage of $\alpha_{PT}$ control is that the probability of making at least one Type I error increases as the number of tests increases. The actual increase in the probability depends, among other factors, on the degree of correlation among the tests. For independent tests, the probability of a Type I error with T tests is $1 - (1 - \alpha)^T$, whereas for nonindependent tests (e.g., all pairwise comparisons, multiple path coefficients in structural equation modeling), the probability of a Type I error is less than $1 - (1 - \alpha)^T$. In general, the more tests that a researcher conducts in his or her experiment, the more likely it is that one (or more) will be significant simply by chance.

### Familywise Error Rate

The familywise error rate ($\alpha_{FW}$) is defined as the probability of falsely rejecting one or more hypotheses in a family of hypotheses. Controlling $\alpha_{FW}$ is recommended when some effects are likely to be nonsignificant; when the researcher is prepared to perform many tests of significance in order to find a significant result; when the researcher's analysis is exploratory, yet he or she still wants to be confident that a significant result is real; and when replication of the experiment is unlikely.

Although many multiple comparison procedures purport to control $\alpha_{FW}$, procedures are said to provide strong $\alpha_{FW}$ control if $\alpha_{FW}$ is maintained at approximately $\alpha$ when all population means are equal (complete null) and when the complete null

is not true but multiple subsets of the population means are equal (partial nulls). Procedures that control $\alpha_{FW}$ for the complete null, but not for partial nulls, provide weak $\alpha_{FW}$ control.

The main advantage of $\alpha_{FW}$ control is that the probability of making a Type I error does not increase with the number of comparisons conducted in the experiment. One of the main disadvantages of procedures that control $\alpha_{FW}$ is that $\alpha_T$ decreases, often substantially, as the number of tests increases. Therefore, procedures that control $\alpha_{FW}$ have reduced power for detecting treatment effects when there are many comparisons, increasing the potential for inconsistent results between experiments.

#### False Discovery Rate

The false discovery rate represents a compromise between strict $\alpha_{FW}$ control and liberal $\alpha_{PT}$ control. Specifically, the false discovery rate ($\alpha_{FDR}$) is defined as the expected ratio ($Q$) of the number of erroneous rejections ($V$) to the total number of rejections ($R = V + S$), where $S$ represents the number of true rejections. Therefore, $E(Q) = E(V/[V+S]) = E(V/R)$.

If all null hypotheses are true, $\alpha_{FDR} = \alpha_{FW}$. On the other hand, if some null hypotheses are false, $\alpha_{FDR} \leq \alpha_{FW}$, resulting in weak control of $\alpha_{FM}$. As a result, any procedure that controls $\alpha_{FW}$ also controls $\alpha_{FDR}$, but procedures that control $\alpha_{FDR}$ can be much more powerful than those that control $\alpha_{FW}$, especially when a large number of tests are performed, and do not entirely dismiss the multiplicity issue (as with $\alpha_{PT}$ control). Although some researchers recommend exclusive use of $\alpha_{FDR}$ control, it is often recommended that $\alpha_{FDR}$ control be reserved for exploratory research, nonsimultaneous inference (e.g., if one had multiple dependent variables and separate inferences would be made for each), and very large family sizes (e.g., as in an investigation of potential activation of thousands of brain voxels in functional magnetic resonance imaging).

## Multiple Comparison Procedures

This section introduces some of the multiple comparison procedures that are available for controlling $\alpha_{FW}$ and $\alpha_{FDR}$. Recall that no multiple comparison procedure is necessary for controlling $\alpha_{PT}$ as the $\alpha$ level for each test is equal to the global $\alpha$ level and can therefore be used seamlessly with any test statistic. It is important to note that the procedures introduced here are only a small subset of the procedures that are available, and for the procedures that are presented, specific details are not provided. Please see specific sections of the encyclopedia for details on these procedures.

### *Familywise Error Controlling Procedures for Any Multiple Testing Environment*

#### Bonferroni

This simple-to-use procedure sets $\alpha_T = \alpha/T$, where $T$ represents the number of tests being performed. The important assumption of the Bonferroni procedure is that the tests being conducted are independent. When this assumption is violated (and it commonly is), the procedure will be too conservative.

#### Dunn-Šidák

The Dunn-Šidák procedure is a more powerful version of the original Bonferroni procedure. With the Dunn-Šidák procedure, $\alpha_T = 1 - (1 - \alpha)^{1/T}$.

#### Holm

Sture Holm proposed a sequential modification of the original Bonferroni procedure that can be substantially more powerful than the Bonferroni or Dunn-Šidák procedures.

#### Hochberg

Yosef Hochberg proposed a modified segue to Bonferroni procedure that combined Simes's inequality with Holm's testing procedure to create a multiple comparison procedure that can be more powerful and simpler than the Holm procedure.

### *Familywise Error Controlling Procedures for Pairwise Multiple Comparison Tests*

#### Tukey

John Tukey proposed the honestly significant difference procedure, which accounts for dependencies among the pairwise comparisons and is

maximally powerful for simultaneous testing of all pairwise comparisons.

#### Hayter

Anthony Hayter proposed a modification to Fisher's least significant difference procedure that would provide strong control over $\alpha_{FW}$.

#### REGWQ

Ryan proposed a modification to the Newman–Keuls procedure that ensures that $\alpha_{FW}$ is maintained at $\alpha$, even in the presence of multiple partial null hypotheses. Ryan's original procedure became known as the REGWQ after modifications to the procedure by Einot, Gabriel, and Welsch.

### *False Discovery Rate–Controlling Procedures for Any Multiple Testing Environment*

#### FDR-BH

Yoav Benjamini and Hochberg proposed the original procedure for controlling the $\alpha_{FDR}$, which is a sequential modified Bonferroni procedure.

#### FDR-BY

Benjamini and Daniel Yekutieli proposed a modified $\alpha_{FDR}$ controlling procedure that would maintain $\alpha_{FDR} = \alpha$ under any dependency structure among the tests.

When multiple tests of significance are performed, the best form of Type I error control depends on the nature and goals of the research. The practical implications associated with the statistical conclusions of the research should take precedence when selecting a form of Type I error control.

*Robert A. Cribbie*

***See also*** Bonferroni Procedure; Coefficient Alpha; Error Rates; False Positive; Holm's Sequential Bonferroni Procedure; Honestly Significant Difference (HSD) Test; Mean Comparisons; Newman–Keuls Test and Tukey Test; Pairwise Comparisons; Post Hoc Comparisons; Scheffé Test; Significance Level, Concept of; Tukey's Honestly Significant Difference (HSD); Type I Error

#### Further Readings

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B, 57,* 289–300.

Hancock, G. R., & Klockars, A. J. (1996). The quest for $\alpha$: Developments in multiple comparison procedures in the quarter century since Games (1971). *Review of Educational Research, 66,* 269–306.

Hochberg, Y., & Tamhane, A. C. (1987). *Multiple comparison procedures.* New York: Wiley.

Hsu, J. C. (1996). *Multiple comparisons: Theory and methods.* New York: Chapman & Hall/CRC.

Toothaker, L. (1993). *Multiple comparison procedures.* Newbury Park, CA: Sage.

Westfall, P. H., Tobia, R. D., Rom, D., Wolfinger, R. D., & Hochberg, Y. (1999). *Multiple comparisons and multiple tests using the SAS system.* Cary, NC: SAS Institute.

## Multiple Regression

Multiple regression is a general and flexible statistical method for analyzing associations between two or more independent variables and a single dependent variable. As a general statistical technique, multiple regression can be employed to predict values of a particular variable based on knowledge of its association with known values of other variables, and it can be used to test scientific hypotheses about whether and to what extent certain independent variables explain variation in a dependent variable of interest. As a flexible statistical method, multiple regression can be used to test associations among continuous as well as categorical variables, and it can be used to test associations between individual independent variables and a dependent variable, as well as interactions among multiple independent variables and a dependent variable. In this entry, different approaches to the use of multiple regression are presented, along with explanations of the more commonly used statistics in multiple regression, methods of conduction multiple regression analysis, and the assumptions of multiple regression.

### Approaches to Using Multiple Regression

#### *Prediction*

One common application of multiple regression is for predicting values of a particular dependent

variable based on knowledge of its association with certain independent variables. In this context, the independent variables are commonly referred to as *predictor* variables and the dependent variable is characterized as the *criterion* variable. In applied settings, it is often desirable for one to be able to predict a score on a criterion variable by using information that is available in certain predictor variables. For example, in the life insurance industry, actuarial scientists use complex regression models to predict, on the basis of certain predictor variables, how long a person will live. In scholastic settings, college and university admissions offices will use predictors such as high school grade point average (GPA) and ACT scores to predict an applicant's college GPA, even before he or she has entered the university.

Multiple regression is most commonly used to predict values of a criterion variable based on *linear* associations with predictor variables. A brief example using simple regression easily illustrates how this works. Assume that a horticulturist developed a new hybrid maple tree that grows exactly 2 feet for every year that it is alive. If the height of the tree was the criterion variable and the age of the tree was the predictor variable, one could accurately describe the relationship between the age and height of the tree with the formula for a straight line, which is also the formula for a simple regression equation:

$$Y = bX + a,$$

where $Y$ is the value of the dependent, or criterion, variable; $X$ is the value of the independent, or predictor, variable; b is a regression coefficient that describes the slope of the line; and a is the $Y$ intercept. The $Y$ intercept is the value of $Y$ when $X$ is 0. Returning to the hybrid tree example, the exact relationship between the tree's age and height could be described as follows:

$$\text{height} = 2(\text{age in years}) + 0$$

Notice that the $Y$ intercept is 0 in this case because at 0 years of age, the tree has 0 height. At that point, it is just seed in the ground. It is clear how knowledge of the relationship between the tree's age and height could be used to easily predict the height of any given tree by just knowing what its age is. A 5-year-old tree will be 10 feet tall, an 8-year-old tree will be 16 feet tall, and so on.

At this point two important issues must be considered. First, virtually any time one is working with variables from people, animals, plants, and so forth, there are no perfect linear associations. Sometimes students with high ACT scores do poorly in college whereas some students with low ACT scores do well in college. This shows how there can always be some *error* when one uses regression to predict values on a criterion variable. The stronger the association between the predictor and criterion variable, the less error there will be in that prediction. Accordingly, regression is based on the *line of best fit*, which is simply the line that will best describe or capture the relationship between $X$ and $Y$ by minimizing the extent to which any data points fall off that line.

A college admissions committee wants to be able to predict the graduating GPA of the students whom they admit. The ACT score is useful for this, but as noted above, it does not have a perfect association with college GPA, so there is some error in that prediction. This is where multiple regression becomes very useful. By taking into account the association of additional predictor variables with college GPA, one can further minimize the error in predicting college GPA. For example, the admissions committee might also collect information on high school GPA and use that in conjunction with the ACT score to predict college GPA. In this case, the regression equation would be

$$Y' = b_1X_1 + b_2X_2 + a$$

where $Y'$ is the predicted value of $Y$ (college GPA), $X_1$ and $X_2$ are the values of the predictor variables (ACT score and high school GPA), $b_1$ and $b_2$ are the regression coefficients by which $X_1$ and $X_2$ are multiplied to get $Y'$, and $a$ is the intercept (i.e., value of $Y$ when $X_1$ and $X_2$ are both 0). In this particular case, the intercept serves only an arithmetic function as it has no practical interpretability because having a high school GPA of 0 and an ACT score of 0 is meaningless.

### *Explanation*

In social scientific contexts, multiple regression is rarely used to predict unknown values on

a criterion variable. In social scientific research, values of the independent and dependent variables are almost always known. In such cases multiple regression is used to test whether and to what extent the independent variables explain the dependent variable. Most often the researcher has theories and hypotheses that specify causal relations among the independent variables and the dependent variable. Multiple regression is a useful tool for testing such hypotheses. For example, an economist is interested in testing a hypothesis about the determinants of workers' salaries. The model being tested could be depicted as follows:

$$\text{family of origin SES} \rightarrow \text{education} \rightarrow \text{salary},$$

where SES stands for socioeconomic status.

In this simple model, the economist hypothesizes that the SES of one's family of origin will influence how much formal education one acquires, which in turn will predict one's salary. If the economist collected data on these three variables from a sample of workers, the hypotheses could be tested with a multiple regression model that is comparable to the one presented previously in the college GPA example:

$$Y = b_1X_1 + b_2X_2 + a$$

In this case, what was $Y'$ is now $Y$ because the value of $Y$ is known. It is useful to deconstruct the components of this equation to show how they can be used to test various aspects of the economist's model.

In the equation above, $b_1$ and $b_2$ are the *partial regression coefficients*. They are the weights by which one multiplies the value of $X_1$ and $X_2$ when all variables are in the equation. In other words, they represent the expected change in $Y$, per unit of $X$ when all other variables are accounted for, or held constant. Computationally, the values of $b_1$ and $b_2$ can be determined easily by simply knowing the zero-order correlations among all possible pairwise combinations of $Y$, $X_1$, and $X_2$, as well as the standard deviations of the three variables:

$$b_1 = \frac{r_{YX_1} - r_{YX_2}r_{X_1X_2}}{1 - r^2_{X_1X_2}} \cdot \frac{s_Y}{s_{X_1}};$$
$$b_2 = \frac{r_{YX_2} - r_{YX_1}r_{X_1X_2}}{1 - r^2_{X_1X_2}} \cdot \frac{s_Y}{s_{X_2}},$$

where $r_{YX_1}$ is the Pearson correlation between $Y$ and $X_1$, $r_{X_1X_2}$ is the Pearson correlation between $X_1$ and $X_2$, and so on, and $s_Y$ is the standard deviation of variable $Y$, $s_{X_1}$ is the standard deviation of $X_1$, and so on. The partial regression coefficients are also referred to as *unstandardized* regression coefficients because they represent the value by which one would multiply the raw $X_1$ or $X_2$ score in order to arrive at $Y$. In the salary example, these coefficients could look something like this:

$$Y = 745.67X_1 + 104.36X_2 + 11{,}325.$$

This means that subjects' annual salaries are best described by an equation whereby their family of origin SES is multiplied by 745.67, their years of formal education are multiplied by 104.36, and these products are added to 11,325. Notice how the regression coefficient for SES is much larger than that for years of formal education. Although it might be tempting to assume that family of origin SES is weighted more heavily than years of formal education, this would not necessarily be correct. The magnitude of an unstandardized regression coefficient is strongly influenced by the units of measurement used to assess the independent variable with which it is associated. In this example, assume that SES is measured on a 5-point scale (Levels 1–5) and years of formal education, at least in the sample, runs from 7 to 20. These differing scale ranges have a profound effect on the magnitude of each regression coefficient, rendering them incomparable.

However, it is often the case that researchers want to understand the relative importance of each independent variable for explaining variation in the dependent variable. In other words, which is the more powerful determinant of people's salaries, their education or their family of origin's socioeconomic status? This question can be evaluated by examining the *standardized regression coefficient*, or $\beta$. Computationally, $\beta$ can be determined by the following formulas:

$$\beta_1 = \frac{r_{YX_1} - r_{YX_2}r_{X_1X_2}}{1 - r^2_{X_1X_2}}; \beta_2 = \frac{r_{YX_2} - r_{YX_1}r_{X_1X_2}}{1 - r^2_{X_1X_2}}.$$

The components of these formulas are identical to those for the unstandardized regression coefficients, but they lack multiplication by the ratio of standard deviations of $Y$ and $X_1$ and $X_2$. Incidentally, one can

easily convert $\beta$ to $b$ with the following formulas, which illustrate their relationship:

$$b_1 = \beta_1 \frac{s_Y}{s_1} \qquad b_2 = \beta_2 \frac{s_Y}{s_2},$$

where $s_Y$ is the standard deviation of variable $Y$, and so on.

Standardized regression coefficients can be thought of as the weight by which one would multiply a standardized score (or $z$ score) for each independent variable in order to arrive at the $z$ score for the dependent variable. Because $z$ scores essentially equate all variables on the same scale, researchers are inclined to make comparisons about the relative impact of each independent variable by comparing their associated standardized regression coefficients, sometimes called *beta weights*.

In the economist's hypothesized model of workers' salaries, there are several subhypotheses or research questions that can be evaluated. For example, the model presumes that both family of origin SES and education will exert a causal influence on annual salary. One can get a sense of which variable has a greater impact on salary by comparing their beta weights. However, it is also important to ask whether either of the independent variables is a significant predictor of salary. In effect, these tests ask whether each independent variable explains a statistically significant portion of the variance in the dependent variable, independent of that explained by the other independent variable(s) also in the regression equation. This can be accomplished by dividing the $\beta$ by its standard error ($SE_\beta$). This ratio is distributed as $t$ with degrees of freedom $= n - k - 1$, where $n$ is the sample size and $k$ is the number of independent variables in the regression analysis. Stated more formally,

$$t = \frac{\beta}{SE_\beta}.$$

If this ratio is significant, that implies that the particular independent variable uniquely explains a statistically significant portion of the variance in the dependent variable. These $t$ tests of the statistical significance of each independent variable are routinely provided by computer programs that conduct multiple regression analyses. They play an important role in testing hypotheses about the role of each independent variable in explaining the dependent variable.

In addition to concerns about the statistical significance and relative importance of each independent variable for explaining the dependent variable, it is important to understand the collective function of the independent variables for explaining the dependent variable. In this case, the question is whether the independent variables collectively explain a significant portion of the variance in scores on the dependent variable. This question is evaluated with the *multiple correlation coefficient*. Just as a simple bivariate correlation is represented by $r$, the multiple correlation coefficient is represented by $R$. In most contexts, data analysts prefer to use $R^2$ to understand the association between the independent variables and the dependent variable. This is because the squared multiple correlation coefficient can be thought of as the percentage of variance in the dependent variable that is collectively explained by the independent variables. So, an $R^2$ value of .65 implies that 65% of the variance in the dependent variable is explained by the combination of independent variables. In the case of two independent variables, the formula for the squared multiple correlation coefficient can be explained as a function of the various pairwise correlations among the independent and dependent variables:

$$R_2 = \frac{r^2_{YX_1} + r^2_{YX_2} - 2r_{YX_I} r_{YX_2} r_{X_1X_2}}{1 - r^2_{X_1X_2}}.$$

In cases with more than two independent variables, this formula becomes much more complex, requiring the use of matrix algebra. In such cases, calculation of $R^2$ is ordinarily left to a computer.

The question of whether the collection of independent variables explains a statistically significant amount of variance in the dependent variable can be approached by testing the multiple correlation coefficient for statistical significance. The test can be carried out by the following formula:

$$F = \frac{R^2(n - k - 1)}{(1 - R^2)k}.$$

This test is distributed as $F$ with $df = k$ in the numerator and $n - k - 1$ in the denominator,

where $n$ is the sample size and $k$ is the number of independent variables.

Two important features of the test for significance of the multiple correlation coefficient require discussion. First, notice how the sample size, $n$, appears in the numerator. This implies that all other things held constant, the larger the sample size, the larger the $F$ ratio will be. That means that the statistical significance of the multiple correlation coefficient is more probable as the sample size increases. Second, the amount of variation in the dependent variable that is not explained by the independent variables, indexed by $1 - R^2$ (this is called error or *residual variance*), is multiplied by the number of independent variables, $k$. This implies that all other things held equal, the larger the number of independent variables, the larger the denominator, and hence the smaller the $F$ ratio. This illustrates how there is something of a penalty for using a lot of independent variables in a regression analysis. When trying to explain scores on a dependent variable, such as salary, it might be tempting to use a large number of predictors so as to take into account as many possible causal factors as possible. However, as this formula shows, this significance test favors parsimonious models that use only a few key predictor variables.

## Methods of Variable Entry

Computer programs used for multiple regression provide several options for the order of entry of each independent variable into the regression equation. The order of entry can make a difference in the results obtained and therefore becomes an important analytic consideration. In *hierarchical* regression, the data analyst specifies a particular order of entry of the independent variables, usually in separate steps for each. Although there are multiple possible logics by which one would specify a particular order of entry, perhaps the most common is that of causal priority. Ordinarily, one would enter independent variables in order from the most distal to the most proximal causes. In the previous example of the workers' salaries, a hierarchical regression analysis would enter family of origin SES into the equation first, followed by years of formal education. As a general rule, in hierarchical entry, an independent variable entered into the equation later should never be the cause of an independent variable entered into the equation earlier. Naturally, hierarchical regression analysis is facilitated by having a priori theories and hypotheses that specify a particular order of causal priority.

Another method of entry that is based purely on empirical rather than theoretical considerations is *stepwise* entry. In this case, the data analyst specifies the full compliment of potential independent variables to the computer program and allows it to enter or not enter these variables into the regression equation, based on the strength of their unique association with the dependent variable. The program keeps entering independent variables up to the point at which addition of any further variables would no longer explain any statistically significant increment of variance in the dependent variable. Stepwise analysis is often used when the researcher has a large collection of independent variables and little theory to explain or guide their ordering or even their role in explaining the dependent variable. Because stepwise regression analysis capitalizes on chance and relies on a post hoc rationale, its use is often discouraged in social scientific contexts.

## Assumptions of Multiple Regression

Multiple regression is most appropriately used as a data analytic tool when certain assumptions about the data are met. First, the data should be collected through independent random sampling. Independent means that the data provided by one participant must be entirely unrelated to the data provided by another participant. Cases in which husbands and wives, college roommates, or doctors and their patients both provide data would violate this assumption. Second, multiple regression analysis assumes that there are linear relationships between the independent variables and the dependent variable. When this is not the case, a more complex version of multiple regression known as *nonlinear regression* must be employed. A third assumption of multiple regression is that at each possible value of each independent variable, the dependent variable must be normally distributed. However, multiple regression is reasonably robust in the case of modest violations of this assumption. Finally, for each possible value of each independent variable, the variance of the residuals or errors in predicting $Y$ (i.e., $Y' - Y$) must be

consistent. This is known as the homoscedasticity assumption. Returning to the workers' salaries example, it would be important that at each level of family-of-origin SES (Levels 1–5), the degree of error in predicting workers' salaries was comparable. If the salary predicted by the regression equation was within ±$5,000 for everyone at Level 1 SES, but it was within ±$36,000 for everyone at Level 3 SES, the homoscedasticity assumption would be violated because there is far greater variability in residuals at the higher compared with lower SES levels. When this happens, the validity of significance tests in multiple regression becomes compromised.

*Chris Segrin*

***See also*** Bivariate Regression; Coefficients of Correlation, Alienation, and Determination; Correlation; Logistic Regression; Pearson Product-Moment Correlation Coefficient; Regression Coefficient

**Further Readings**

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.

Alison, P. D. (1999). *Multiple regression: A primer*. Thousand Oaks, CA: Sage.

Berry, W. D. (1993). *Understanding regression assumptions*. Thousand Oaks, CA: Sage.

Cohen, J., Cohen, P., West, S., & Aiken, L. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research* (3rd ed.). New York: Wadsworth.

## Multiple Treatment Interference

Multiple treatment interference is a threat to the internal validity of a group design. A problem occurs when participants in one group have received all or some of a treatment in addition to the one assigned as part of an experimental or quasi-experimental design. In these situations, the researcher cannot determine what, if any, influence on the outcome is associated with the nominal treatment and what variance is associated with some other treatment or condition. In terms of independent and dependent variable designations, multiple treatment interference occurs when participants were meant to be assigned to one level of the independent variable (e.g., a certain group with a researcher assigned condition) but were functionally at a different level of the variable (e.g., they received some of the treatment meant for a comparison group). Consequently, valid conclusions about cause and effect are difficult to make.

There are several situations that can result in multiple treatment interference, and they can occur in either experimental designs (which have random assignment of participants to groups or levels of the independent variable) or quasi-experimental designs (which do not have random assignment to groups). One situation might find one or more participants in one group receiving accidentally, in addition to their designated treatment, the treatment meant for a second group. This can happen administratively in medicine studies, for example, if subjects receive both the drug they are meant to receive and, accidentally, are also given the drug meant for a comparison group. If benefits are found in both groups or in the group meant to receive a placebo (for example), it is unclear whether effects are due to the experimental drug, the placebo, or a combination of the two. The ability to isolate the effects of the experimental drug or (more generally in research design) the independent variable on the outcome variable is the strength of a good research design, and consequently, strong research designs attempt to avoid the threat of multiple treatment interference. A second situation involving multiple treatment interference is more common, especially in the social sciences. Imagine an educational researcher interested in the effects of a new method of reading instruction. The researcher has arranged for one elementary teacher in a school building to use the experimental approach and another elementary teacher to use the traditional method. As is typically the case in educational research, random assignment to the two classrooms is not possible. Scores on a reading test are collected from both classrooms as part of a pre-post test design. The design looks likes this:

Experimental Group:
Pretest → 12 weeks of instruction → Posttest

Comparison Group:
Pretest → 12 weeks of instruction → Posttest

If the study were conducted as planned, comparisons of posttest means for the two groups, perhaps after controlling for initial differences found at the time of the pretest, would provide fairly valid evidence of the comparative effectiveness of the new method. The conclusion is predicated, though, on the assumption that students in the traditional classroom were not exposed to the experimental method. Often, in the real world, it is difficult to keep participants in the control or comparison group free from the "contamination" of the experimental treatment. In the case of this example, the teacher in the comparison group may have used some of the techniques or strategies included in the experimental approach. He may have done this inadvertently, or intentionally, deciding that ethically he should use the best methods he knows. Contamination might also have been caused by the students' partial exposure to the new instructional approach in some circumstance outside of the classroom—a family member in the other classroom may have brought homework home and shared it with a student, for example.

Both of these examples describe instances of multiple treatment interference. Because of participants' exposure to more than only the intended treatment, it becomes difficult for researchers to establish relationships between well-defined variables. When a researcher believes that multiple treatments might be effective or wishes to investigate the effect of multiple treatments, however, a design can be applied that controls the combinations of treatments and investigates the consequences of multiple treatments. Drug studies sometimes are interested in identifying the benefits or disadvantages of various interactions or combinations of treatments, for example. A study interested in exploring the effects of multiple treatments might look like this (assuming random assignment):

Group 1:
Treatment 1 → Measure outcome

Group 2:
Treatment 2 → Measure outcome

Group 3:
Treatment 1, followed by Treatment 2 → Measure outcome

Group 4:
Treatment 2, followed by Treatment 1 → Measure outcome

A statistical comparison of the four groups' outcome means would identify the optimum treatment—Treatment 1, Treatment 2, or a particular sequence of both treatments.

*Bruce Frey*

***See also*** Experimental Design; Quasi-Experimental Design

**Further Readings**

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton Mifflin.

# Multitrait–Multimethod Matrix

The multitrait–multimethod (MTMM) matrix contains the correlations between variables when each variable represents a trait–method unit, that is, the measurement of a trait (e.g., extroversion, neuroticism) by a specific method (e.g., self-report, peer report). In order to obtain the matrix, each trait has to be measured by the same set of methods. This makes it possible to arrange the correlations in such a way that the correlations between different traits measured by the same method can be separated from the correlations between different traits measured by different methods. The MTMM matrix was recommended by Donald T. Campbell and Donald W. Fiske as a means of measuring the convergent and discriminant validity. This entry discusses the structure, evaluation, and analysis approaches of the MTMM matrix.

## Structure of the MTMM Matrix

Table 1 shows a prototypical MTMM matrix for three traits measured by three methods. An

**Table 1** Multitrait–Multimethod Matrix for Three Traits Measured by Three Methods

| | | Method 1 | | | Method 2 | | | Method 3 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Traits | Trait 1 | Trait 2 | Trait 3 | Trait 1 | Trait 2 | Trait 3 | Trait 1 | Trait 2 | Trait 3 |
| Method 1 | Trait 1 | (.90) | | | | | | | | |
| | Trait 2 | .10 | (.90) | | | | | | | |
| | Trait 3 | .10 | .10 | (.90) | | | | | | |
| Method 2 | Trait 1 | **.50** | .12 | .12 | (.80) | | | | | |
| | Trait 2 | .14 | **.50** | .12 | .20 | (.80) | | | | |
| | Trait 3 | .14 | .14 | **.50** | .20 | .20 | (.80) | | | |
| Method 3 | Trait 1 | **.50** | .05 | .05 | **.40** | .03 | .03 | (.85) | | |
| | Trait 2 | .04 | **.50** | .05 | .02 | **.40** | .03 | .30 | (.85) | |
| | Trait 3 | .04 | .04 | **.50** | .02 | .02 | **.40** | .30 | .30 | (.85) |

*Notes:* The correlations are artificial. Reliabilities are in parentheses. Heterotrait–monomethod correlations are in the gray subdiagonals. Heterotrait–heteromethod correlations are enclosed by a broken line. Monotrait–heteromethod correlations in the convergent validity diagonals are in bold type.

MTMM matrix consists of two major parts: monomethod blocks and heteromethod blocks.

### *Monomethod Blocks*

The monomethod blocks contain the correlations between variables that belong to the same method. In Table 1 there are three monomethod blocks, one for each method. Each monomethod block consists of two parts. The first part (reliability diagonals) contains the reliabilities of the measures. The second part (the heterotrait–monomethod triangles) include the correlations between different traits that are measured by the same methods. The reliabilities can be considered as monotrait–monomethod correlations.

### *Heteromethod Blocks*

The heteromethod blocks comprise the correlations between traits that were measured by different methods. Table 1 contains three heteromethod blocks, one for each combination of the three methods. A heteromethod block consists of two parts. The validity diagonal (monotrait–heteromethod correlations) contains the correlations of the same traits measured by different methods. The heterotrait–heteromethod triangles cover the correlations of different traits measured by different methods.

## Criteria for Evaluating the MTMM Matrix

Campbell and Fiske described four properties an MTMM matrix should show when convergent and discriminant validity is present:

1. The correlations in the validity diagonals (monotrait–heteromethod correlations) should be significantly different from 0 and they should be large. These correlations indicate convergent validity.
2. The heterotrait–heteromethod correlations should be smaller than the monotrait–heteromethod correlations (discriminant validity).
3. The heterotrait–monomethod correlations should be smaller than the montrait–heteromethod correlations (discriminant validity).
4. The same pattern of trait intercorrelations should be shown in all heterotrait triangles in

the monotrait as well as in the heteromethod blocks (discriminant validity).

### *Limitations of These Rules*

These four requirements have been developed by Campbell and Fiske because they are easy-to-apply rules for evaluating an MTMM matrix with respect to its convergent and discriminant validity. They are, however, restricted in several ways. The application of these criteria is difficult if the different measures differ in their reliabilities. In this case the correlations, which are correlations of observed variables, can be distorted by measurement error in different ways, and differences between correlations could only be due to differences in reliabilities. Moreover, there is no statistical test of whether these criteria are fulfilled in a specific application. Finally, the MTMM matrix is not explained by a statistical model allowing the separation of different sources of variance that are due to trait, method, and error influences. Modern psychometric approaches complete these criteria and circumvent some of these problems.

## Modern Psychometric Approaches for Analyzing the MTMM Matrix

Many statistical approaches, such as analysis of variance and generalizability theory, multilevel modeling, and item response theory, have been applied to analyze MTMM data sets. Among all methods, direct product models and models of confirmatory factor analyses have been the most often applied and influential approaches.

### *Direct Product Models*

The basic idea of direct product models is that each correlation of an MTMM matrix is assumed to be a product of two correlations: a correlation between traits and a correlation between methods. For each combination of traits, there is a correlation indicating discriminant validity, and for each combination of methods, there is a correlation indicating the degree of convergent validity. For example, the heterotrait–heteromethod correlation $Cor(T_1M_1, T_2M_2)$ between a trait $T_1$ (measured by a method $M_1$) and a trait $T_2$ (measured by a method $M_2$) is the product of the correlations $Cor(T_1, T_2)$ and $Cor(M_1, M_2)$:

$$Cor(T_1M_1, T_2M_2) = Cor(T_1, T_2) \times Cor(M_1, M_2).$$

If the two traits are measured by the same method (e.g., $M_1$), the method intercorrelation is 1 ($Cor[M_1, M_1] = 1$), and the observed correlation equals the correlation between the two traits:

$$Cor(T_1M_1, T_2M_1) = Cor(T_1, T_2).$$

If the two traits, however, are measured by different methods, the correlation of the traits is attenuated by the correlation of the methods. Hence, the smaller the convergent validity, the smaller are the expected correlations between the traits.

The four properties of the MTMM matrix proposed by Campbell and Fiske can be evaluated by the correlations of the direct product model:

1. The correlations between two methods (convergent validity) should be large.
2. The second property is always fulfilled when the correlation between two traits is smaller than 1, because the direct product model always implies, in this case,
$$Cor(T_1, T_1) \times Cor(M_1, M_2) > Cor(T_1, T_2) \times Cor(M_1, M_2).$$
3. The third property is satisfied when the correlations between methods are larger than the correlations between traits because $Cor(T_1, T_1) = Cor(M_1, M_1) = 1$, and in this case,
$$Cor(T_1, T_1) \times Cor(M_1, M_2) > Cor(T_1, T_2) \times Cor(M_1, M_1).$$
4. The fourth requirement is always fulfilled if the direct product model holds because the trait intercorrelations in all heteromethod blocks are weighted by the same method correlation. This makes sure that the ratio of two trait correlations is the same for all mono-method and heteromethod blocks.

The direct product model has been extended by Michael Browne to the *composite direct product model* by the consideration of measurement error. Direct product models are reasonable models for analyzing MTMM matrices. They are, however,

also limited in at least two respects. First, they do not allow the decomposition of variance in components due to trait, method, and error influences. Second, they presuppose that the correlations between the traits do not differ between the different monomethod blocks.

### *Models of Confirmatory Factor Analysis*

During recent years many different MTMM models of confirmatory factor analysis (CFA) have been developed. Whereas the first models were developed for decomposing the classical MTMM matrix that is characterized by a single indicator for each trait–method unit, more recently formulated models consider multiple indicators for each trait–method unit.

#### Single Indicator Models

Single indicator models of CFA decompose the observed variables into different components representing trait, method, and error influences. Moreover, they differ in the assumptions they make concerning the homogeneity of method and trait effects, as well as admissible correlations between trait and method factors. Keith Widaman, for example, describes a taxonomy of 16 MTMM-CFA models by combining four different types of trait structures (no trait factor, general trait factor, several orthogonal trait factors, several oblique trait factors) with four types of method structures (no method factor, general method factor, several orthogonal method factors, several oblique method factors).

In the most general single-indicator MTMM model, an observed variable $Y_{jk}$, indicating a trait $i$ measured by a method $k$, is decomposed into a trait factor $T_j$, a method factor $M_k$, and a residual variable $E_{jk}$:

$$Y_{jk} = \alpha_{jk} + \lambda_{\mathrm{T}jk} T_j + \lambda_{\mathrm{M}jk} M_k + E_{jk},$$

where $\alpha_{jk}$ is an intercept, $\lambda_{\mathrm{T}jk}$ is a trait loading, and $\lambda_{\mathrm{M}jk}$ is a method loading. The trait factors are allowed to be correlated. The correlations indicate the degree of discriminant validity. Also, the method factors are allowed to be correlated. These method intercorrelations show whether method effects generalize across methods. This model is called a *correlated-trait–correlated-method* (CTCM) *model* and is depicted in Figure 1.

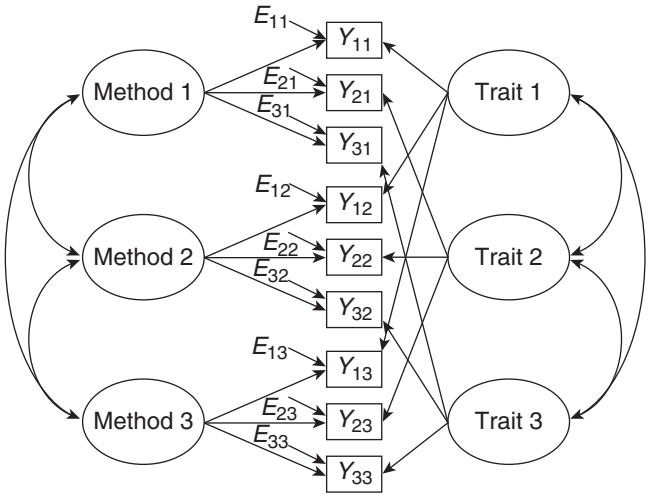



**Figure 1** Correlated-Trait–Correlated-Method Model

The model has several advantages. It allows researchers to decompose the variance of an observed variable into the variance due to trait, method, and error influences. Convergent validity is given when the variances due to the method factors are small. In the ideal case of perfect convergent validity, the variances of all method factors would be 0, meaning that there are no method factors in the model (so-called *correlated trait model*). Perfect discriminant validity could be present if the trait factors are uncorrelated. Method factors could be correlated, indicating that the different methods can be related in a different way. However, aside from these advantages, this model is affected by serious statistical and conceptual problems. One statistical problem is that the model is not generally identified, which means that there are data constellations in which it is not possible to estimate the parameters of the model. For example, if all factor loadings do not differ from each other, the model is not identified. Moreover, applications of this model often show nonadmissible parameter estimates (e.g., negative variances of method factors), and the estimation process often does not converge. From a more conceptual point of view, the model has been criticized because it allows correlations between method factors. These correlations make the interpretation of the correlations between trait factors more difficult. If, for

example, all trait factors are uncorrelated, this would indicate perfect discriminant validity. However, if all method factors are correlated, it is difficult to interpret the uncorrelatedness of the trait factors as perfect discriminant validity because the method factor correlations represent a portion of variance shared by all variables that might be due to a general trait effect.

The major problems of the CTCM model are caused by the correlated method factors. According to Michael Eid, Tanja Lischetzke, and Fridtjof Nussbeck, the problems of the CTCM model can be circumvented by dropping the correlations between the method factors or by dropping one method factor. A CTCM model without correlations between method factors is called a *correlated-trait–uncorrelated-method* (CTUM) model. This model is a special case of the model depicted in Figure 1 but with uncorrelated method factors. This model is reasonable if correlations between method factors are not expected. According to Eid and colleagues, this is the case when interchangeable methods are considered. Interchangeable methods are methods that are randomly chosen from a set of methods. If one considers different raters as different methods, an example of interchangeable raters (methods) is randomly selected students rating their teacher. If one randomly selects three students for each teacher and if one assigns these three students randomly to three rater groups, the three method factors would represent the deviation of individual raters from the expected (mean) rating of the teacher (the trait scores). Because the three raters are interchangeable, correlations between the method factors would not be expected. Hence, applying the CTUM model in the case of interchangeable raters would circumvent the problems of the CTCM model.

The situation, however, is quite different in the case of structurally different methods. An example of structurally different methods is a self-rating, a rating by the parents, and a rating by the teacher. In this case, the three raters are not interchangeable but are structurally different. In this case, the CTUM model is not reasonable as it may not adequately represent the fact that teachers and parents can share a common view that is not shared with the student (correlations of method effects). In this case, dropping one method factor solves the problem of the CTCM model in many cases. This model with one method factor less than the number of methods considered is called the *correlated-trait-correlated-(method – 1) model* (CTC[$M - 1$]). This model is a special case of the model depicted in Figure 1 but with one method factor less. If the first method in Figure 1 is the self-report, the second method is the teacher report, and the third method is the parent report, dropping the first method factor would imply that the three trait factors equal the true-score variables of the self-reports. The self-report method would play the role of the reference method that has to be chosen in this model. Hence, in the CTC($M - 1$) model, the trait factor is completely confounded with the reference method. The method factors have a clear meaning. They indicate the deviation of the true (error-free) other reports from the value predicted by the self-report. A method effect is that (error-free) part of a nonreference method that cannot be predicted by the reference method. The correlations between the two method factors would then indicate that the two other raters (teachers and parents) share a common view of the child that is not shared by the child herself or himself. This model allows contrasting methods, but it does not contain common "method-free" trait factors. It is doubtful that such a common trait factor has a reasonable meaning in the case of structurally different methods.

All single indicator models presented so far assume that the method effects belonging to one method are unidimensional as there is one common method factor for each method. This assumption could be too strong as method effects could be trait specific. Trait-specific method effects are part of the residual in the models presented so far. That means that reliability will be underestimated because a part of the residual is due to method effects and not due to measurement error. Moreover, the models may not fit the data in the case of trait-specific method effects. If the assumption of unidimensional method factors for a method is too strong, the method factors can be dropped and replaced by correlated residuals. For example, if one replaces the method factors in the CTUM model by correlations of residuals belonging to the same methods, one obtains the *correlated-trait–correlated-uniqueness* (CTCU) model. However, in this model the reliabilities are underestimated

because method effects are now part of the residuals. Moreover, the CTUM model does not allow correlations between residuals of different methods. This might be necessary in the case of structurally different methods. Problems that are caused by trait-specific method effects can be appropriately handled in multiple indicator models.

### Multiple Indicator Models

In multiple indicator models, there are several indicators for one trait–method unit. In the less restrictive model, there is one factor for all indicators belonging to the same trait–method unit. The correlations between these factors constitute a latent MTMM matrix. The correlation coefficients of this latent MTMM matrix are not distorted by measurement error and allow a more appropriate application of the Campbell and Fiske criteria for evaluating the MTMM matrix. Multiple indicator models allow the definition of trait-specific method factors and, therefore, the separation of measurement error and method-specific influences in a more appropriate way. Eid and colleagues have shown how different models of CFA can be defined for different types of methods. In the case of interchangeable methods, a multilevel CFA model can be applied that allows the specification of trait-specific method effects. In contrast to the extension of the CTCU model to multiple indicators, the multilevel approach has the advantage that the number of methods (e.g., raters) can differ between targets. In the case of structurally different raters, an extension of the CTC(M − 1) model to multiple indicators can be applied. This model allows a researcher to test specific hypotheses about the generalizability of method effects across traits and methods. In the case of a combination of structurally different and interchangeable methods, a multilevel CTC(M − 1) model would be appropriate.

*Michael Eid*

***See also*** Construct Validity; "Convergent and Discriminant Validation by the Multitrait–Multimethod Matrix"; MBESS; Structural Equation Modeling; Triangulation; Validity of Measurement

### Further Readings


Browne, M. W. (1984). The decomposition of multitrait-multimethod matrices. *British Journal of Mathematical & Statistical Psychology, 37,* 1–21.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Dumenci, L. (2000). Multitrait-multimethod analysis. In S. D. Brown & H. E. A. Tinsley (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 583–611). San Diego, CA: Academic Press.

Eid, M. (2000). A multitrait-multimethod model with minimal assumptions. *Psychometrika, 65,* 241–261.

Eid, M. (2006). Methodological approaches for analyzing multimethod data. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 223–230). Washington, DC: American Psychological Association.

Eid, M., & Diener, E. (2006). *Handbook of multimethod measurement in psychology.* Washington, DC: American Psychological Association.

Eid, M., Nussbeck, F. W., Geiser, C., Cole, D. A., Gollwitzer, M., & Lischetzke, T. (2008). Structural equation modeling of multitrait-multimethod data: Different models for different types of methods. *Psychological Methods, 13,* 230–253.

Kenny, D. A. (1976). An empirical application of confirmatory factor analysis to the multitrait-multimethod matrix. *Journal of Experimental Social Psychology, 12,* 247–252.

Marsh, H. W., & Grayson, D. (1995). Latent variable models of multitrait-multimethod data. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 177–198). Thousands Oaks, CA: Sage.

Shrout, P. E., & Fiske, S. T. (Eds.). (1995). *Personality research, methods, and theory: A festschrift honoring Donald W. Fiske.* Hillsdale, NJ: Lawrence Erlbaum


## Multivalued Treatment Effects

The term *multivalued treatment effects* broadly refers to a collection of population parameters that capture the impact of a given treatment assigned to each observational unit, when this treatment status takes multiple values. In general, treatment levels may be finite or infinite as well as ordinal or cardinal, leading to a large collection of possible treatment effects to be studied in applications. When the treatment effect of interest is the mean

outcome for each treatment level, the resulting population parameter is typically called the *dose–response function* in the statistical literature, regardless of whether the treatment levels are finite or infinite. The analysis of multivalued treatment effects has several distinct features when compared with the analysis of binary treatment effects, including the following: (a) A comparison or control group is not always clearly defined, (b) new parameters of interest arise capturing distinct phenomena such as nonlinearities or tipping points, (c) in most cases correct statistical inferences require the joint estimation of all treatment effects (as opposed to the estimation of each treatment effect at a time), and (d) efficiency gains in statistical inferences may be obtained by exploiting known restrictions among the multivalued treatment effects. This entry discusses the treatment effect model and statistical inference procedures for multivalued treatment effects.

## Treatment Effect Model and Population Parameters

A general statistical treatment effect model with multivalued treatment assignments is easily described in the context of the classical potential outcomes model. This model assumes that each unit $i$ in a population has an underlying collection of potential outcome random variables $\{Y_i = Y_i(t) : t \in \mathcal{T}\}$, where $\mathcal{T}$ denotes the collection of possible treatment assignments. The random variables $Y_i(t)$ are usually called *potential outcomes* because they represent the random outcome that unit $i$ would have under treatment regime $t \in \mathcal{T}$. For each unit $i$ and for any two treatment levels, $t_1$ and $t_2$, it is always possible to define the *individual treatment effect* given by $Y_i(t_1) - Y_i(t_2)$, which may or may not be a degenerate random variable. However, because units are not observed under different treatment regimes simultaneously, such comparisons are not feasible. This idea, known as the *fundamental problem of causal inference,* is formalized in the model by assuming that for each unit $i$ only $(Y_i, T_i)$ is observed, where $Y_i = Y_i(T_i)$ and $T_i \in \mathcal{T}$. In words, for each unit $i$, only the potential outcome for treatment level $T_i = t$ is observed while all other (counterfactual) outcomes are missing. Of course, in most applications, which treatment each unit has taken up is not random and hence further assumptions would be needed to identify the treatment effect of interest.

A binary treatment effect model has $\mathcal{T} = \{0, 1\}$, a finite multivalued treatment effect model has $\mathcal{T} = \{0, 1, \ldots, J\}$ for some positive integer $J$, and a continuous treatment effect model has $\mathcal{T} = [0, 1]$. (Note that the values in $\mathcal{T}$ are ordinal, that is, they may be seen just as normalizations of the underlying real treatment levels in a given application.) Many applications focus on a binary treatment effects model and base the analysis on the comparison of two groups, usually called *treatment group* $(T_i = 1)$ and *control group* $(T_i = 0)$. A multivalued treatment may be collapsed into a binary treatment, but this procedure usually would imply some important loss of information in the analysis. Important phenomena such as nonlinearities, differential effects across treatment levels or tipping points, cannot be captured by a binary treatment effect model.

Typical examples of multivalued treatment effects are comparisons between some characteristic of the distributions of the potential outcomes. Well-known examples are mean and quantile comparisons, although in many applications other features of these distributions may be of interest. For example, assuming, to simplify the discussion, that the random potential outcomes are equal for all units (this holds, for instance, in the context of random sampling), the mean of the potential outcome under treatment regime $t \in \mathcal{T}$ is given by $\mu(t) = \mathrm{E}[Y_i(t)]$. The collection of these means is the so-called dose–response function. Using this estimand, it is possible to construct different multivalued treatment effects of interest, such as pairwise comparisons (e.g., $\mu(t_2) - \mu(t_1)$) or differences in pairwise comparisons, which would capture the idea of nonlinear treatment effects. (In the particular case of binary treatment effects, the only possible pairwise comparison is $\mu(1) - \mu(0)$, which is called the *average treatment effect.*) Using the dose–response function, it is also possible to consider other treatment effects that arise as nonlinear transformations of $\mu(t)$, such as ratios, incremental changes, tipping points, or the maximal treatment effect $\mu^* = max_{t \in \mathcal{T}} \mu(t)$, among many other possibilities. All these multivalued treatment effects are constructed on the basis of

the mean of the potential outcomes, but similar estimands may be considered that are based on quantiles, dispersion measures, or other characteristics of the underlying potential outcome distribution. Conducting valid hypothesis testing about these treatment effects requires in most cases the joint estimation of the underlying multivalued treatment effects.

### Statistical Inference

There exists a vast theoretical literature proposing and analyzing different statistical inference procedures for multivalued treatment effects. This large literature may be characterized in terms of the key identifying assumption underlying the treatment effect model. This key assumption usually takes the form of a (local) independence or orthogonality condition, such as (a) a conditional independence assumption, which assumes that conditional on a set of observable characteristics, selection into treatment is random, or (b) an instrumental variables assumption, which assumes the existence of variables that induce exogenous changes in the treatment assignment. With the use of an identifying assumption (together with other standard model assumptions), it has been shown in the statistical and econometrics literatures that several parametric, semiparametric, and nonparametric procedures allow for optimal joint inference in the context of multivalued treatments. These results are typically obtained with the use of large sample theory and justify (asymptotically) the use of classical statistical inference procedures involving multiple treatment levels.

*Matias D. Cattaneo*

***See also*** Multiple Treatment Interference; Observational Research; Propensity Score Analysis; Selection; Treatment(s)

#### Further Readings

Cattaneo, M. D. (2010). Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics, 155,* 138–154.

Heckman, J. J., & Vytlacil, E. J. (2007). Econometric evaluation of social programs, Part I: Causal models, structural models and econometric policy evaluation. In J. J. Heckman and E. E. Leamer (Eds.), *Handbook of econometrics* (Vol. 6B, pp. 4779–4874). Amsterdam: North-Holland.

Imai, K., & van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association, 99,* 854–866.

Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature, 47,* 5–86.

Rosembaum, P. (2002). *Observational studies.* New York: Springer.

## Multivariate Analysis of Variance (MANOVA)

Multivariate analysis of variance (MANOVA) designs are appropriate when multiple dependent variables are included in the analysis. The dependent variables should represent continuous measures (i.e., interval or ratio data). Dependent variables should be moderately correlated. If there is no correlation at all, MANOVA offers no improvement over an analysis of variance (ANOVA); if the variables are highly correlated, the same variable may be measured more than once. In many MANOVA situations, multiple independent variables, called factors, with multiple levels are included. The independent variables should be categorical (qualitative). Unlike ANOVA procedures that analyze differences across two or more groups on one dependent variable, MANOVA procedures analyze differences across two or more groups on two or more dependent variables. Investigating two or more dependent variables simultaneously is important in various disciplines, ranging from the natural and physical sciences to government and business and to the behavioral and social sciences. Many research questions cannot be answered adequately by an investigation of only one dependent variable because treatments in experimental studies are likely to affect subjects in more than one way. The focus of this entry is on the various types of MANOVA procedures and associated assumptions. The logic of MANOVA and advantages and disadvantages of MANOVA are included.

MANOVA is a special case of the general *linear models.* MANOVA may be represented in a basic linear equation as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbf{Y}$ represents a vector of dependent variables, $\mathbf{X}$ represents

a matrix of independent variables, $\beta$ represents a vector of weighted regression coefficients, and $\varepsilon$ represents a vector of error terms. Calculations for the multivariate procedures are based on matrix algebra, making hand calculations virtually impossible. For example, the null hypothesis for MANOVA states no difference among the population mean vectors. The form of the omnibus null hypothesis is written as $H_0 = \mu_1 = \cdots = \mu_k$. It is important to remember that the means displayed in the null hypothesis represent mean *vectors* for the population, rather than the population means. The complexity of MANOVA calculations requires the use of statistical software for computing.

## Logic of MANOVA

MANOVA procedures evaluate differences in population means on more than one dependent variable across levels of a factor. MANOVA uses a linear combination of the dependent variables to form a new dependent variable that minimizes within-group variance and maximizes between-group differences. The new variable is used in an ANOVA to compare differences among the groups. Use of the newly formed dependent variable in the analysis decreases the Type I error (error of rejecting a true null hypothesis) rate. The linear combination reveals a more complete picture of the characteristic or attribute under study. For example, a social scientist may be interested in the kinds of attitudes that people have toward the environment based on their attitudes about global warming. In such a case, analysis of only one dependent variable (attitude about global warming) is not completely representative of the attitudes that people have toward the environment. Multiple measures, such as attitude toward recycling, willingness to purchase environmentally friendly products, and willingness to conserve water and energy, will give a more holistic view of attitudes toward the environment. In other words, MANOVA analyzes the composite of several variables, rather than analyzing several variables individually.

## Advantages of MANOVA Designs

MANOVA procedures control for experiment-wide error rate, whereas multiple univariate procedures increase the Type I error rate, which can lead to rejection of a true null hypothesis. For example, analysis of group differences on three dependent variables would require three univariate tests. If the alpha level is set at .05, there is a 95% chance of not making a Type I error. The following calculations show how the 95% error rate is compounded with three univariate tests: $(.95)(.95)(.95) = .857$ and $1 - .857 = .143$, or 14.3%, which is an unacceptable error rate. In addition, univariate tests do not account for the intercorrelations among variables, thus risking loss of valuable information. Furthermore, MANOVA decreases the Type II error (error of not rejecting a false null hypothesis) rate by detecting group differences that appear only through the combination of two or more dependent variables.

## Disadvantages of MANOVA Designs

MANOVA procedures are more complex than univariate procedures; thus, outcomes may be ambiguous and difficult to interpret. The power of MANOVA may actually reveal statistically significant differences when multiple univariate tests may not show differences. Statistical power is the probability of rejecting the null hypothesis when the null is false. (Power $= 1 - \beta$.) The difference in outcomes between ANOVA and MANOVA results from the overlapping of the distributions for each of the groups with the dependent variables in separate analyses. In the MANOVA procedure, the linear combination of dependent variables is used for the analysis. Finally, more assumptions are required for MANOVA than for ANOVA.

## Assumptions of MANOVA

The mathematical underpinnings of inferential statistics require that certain statistical assumptions be met. Assumptions for MANOVA designs are (a) multivariate normality, (b) homoscedasticity, (c) linearity, and (d) independence and randomness.

### *Multivariate Normality*

Observations on all dependent variables are multivariately normally distributed for each level within each group and for all linear combinations of the dependent variables. Joint normality in

more than two dimensions is difficult to assess; however, tests for univariate normality on each of the variables are recommended. Univariate normality, a prerequisite to multivariate normality, can be assessed graphically and statistically. For example, a quantile–quantile plot resembling a straight line suggests normality. While normality of the univariate tests does not mean that the data are multivariately normal, such tests are useful in evaluating the assumption. MANOVA is insensitive (robust) to moderate departures from normality for large data sets and in situations in which the violations are due to skewed data rather than outliers.

A scatterplot for pairs of variables for each group can reveal data points located far from the pattern produced by the other observations. Mahalanobis distance (distance of each case from the centroid of all the remaining cases) is used to detect multivariate outliers. Significance of Mahalanobis distance is evaluated as a chi-square statistic. A case may be considered an outlier if its Mahalanobis distance is statistically significant at the $p < .0001$ level. Other graphical techniques, such as box plots and stem-and-leaf plots, may be used to assess univariate normality. Two additional descriptive statistics related to normality are skewness and kurtosis.

#### Skewness

Skewness refers to the symmetry of the distribution about the mean. Statistical values for skewness range from $\pm\infty$. A perfectly symmetrical distribution will yield a value of zero. In a positively skewed distribution, observations cluster to the left of the mean on the normal distribution curve with the right tail on the curve extended with a small number of cases. The opposite is true for a negatively skewed distribution. Observations cluster to the right of the mean on the normal distribution curve with the left tail on the curve extended with a small number of cases. In general, a skewness value of .7 or .8 is cause for concern and suggests that data transformations may be appropriate.

#### Kurtosis

Kurtosis refers to the degree of peakedness or flatness of a sample distribution compared with a normal distribution. Kurtosis values may be positive or negative to indicate a high peak or flatness near the mean, respectively. Values within $\pm 2$ standard deviations from the mean or $\pm 3$ standard deviations from the mean are generally considered within the normal range. A normal distribution has zero kurtosis. In addition to graphical techniques, the Shapiro–Wilk $W$ statistic and the Kolmogorov–Smirnov statistic with Lilliefors significance levels are used to assess normality. Statistically significant $W$ or Kolmogorov–Smirnov test results indicate that the distribution is nonnormal.

### *Homoscedasticity*

The variance and covariance matrices for all dependent variables across groups are assumed to be equal. George Box's $M$ statistic tests the null hypothesis of equality of the observed covariance matrices for the dependent variables for each group. A nonsignificant $F$ value with the alpha level set at .001 from Box's $M$ indicates equality of the covariance matrices. MANOVA procedures can tolerate moderate departures from equal variance–covariance matrices when sample sizes are similar.

### *Linearity*

MANOVA procedures are based on linear combinations of the dependent variables; therefore, it is assumed that linear relationships exist among all pairs of dependent variables and all pairs of covariates across all groups. Consequently, linearity is important for all dependent variable–covariate pairs as well. Linearity may be assessed by examining scatterplots of pairs of dependent variables for each group. The scatterplot displays an elliptical shape to indicate a linear relationship. If both variables are not normally distributed, the assumption of linearity will not hold, and an elliptical shape will not be displayed on the scatterplot. If the linearity assumption is not met, data transformations may be necessary to establish linearity.

### *Independence and Randomness*

Observations are independent of one another. That is, the score for one participant is independent of scores of any other participants for each

variable. Randomness means that the sample was randomly selected from the population of interest.

## Research Questions for MANOVA Designs

Multivariate analyses cover a broad range of statistical procedures. Common questions for which MANOVA procedures are appropriate are as follows: What are the mean differences between two levels of one independent variable for multiple dependent variables? What are the mean differences between or among multiple levels of one independent variable on multiple dependent variables? What are the effects of multiple independent variables on multiple dependent variables? What are the interactions among independent variables on one dependent variable or on a combination of dependent variables? What are the mean differences between or among groups when repeated measures are used in a MANOVA design? What are the effects of multiple levels of an independent variable on multiple dependent variables when effects of concomitant variables are removed from the analysis? What is the amount of shared variance among a set of variables when variables are grouped around a common theme? What are the relationships among variables that may be useful for predicting group membership? These sample questions provide a general sense of the broad range of questions that may be answered with MANOVA procedures.

## Types of MANOVA Designs

### *Hotelling's* $T^2$

Problems for MANOVA can be structured in different ways. For example, a researcher may wish to examine the difference between males and females on number of vehicle accidents in the past 5 years and years of driving experience. In this case, the researcher has one dichotomous independent variable (gender) and two dependent variables (number of accidents and years of driving experience). The problem is to determine the difference between the weighted sample mean vectors (centroids) of a multivariate data set. This form of MANOVA is known as the multivariate analog to the Student's $t$ test, and it is referred to as Hotelling's $T^2$ statistic, named after Harold Hotelling for his work on the multivariate $T^2$ distribution. Calculation of $T^2$ is based on the combination of two Student's $t$ ratios and their pooled estimate of correlation. The resulting $T^2$ is converted into an $F$ statistic and distributed as an $F$ distribution.

### *One-Way MANOVA*

Another variation of the MANOVA procedure is useful for investigating the effects of one multilevel independent variable (factor) on two or more dependent variables. An investigation of differences in mathematics achievement and motivation for students assigned to three different teaching methods is such a situation. For this problem, the researcher has one multilevel factor (teaching method with three levels) and two dependent variables (mathematics test scores and scores on a motivation scale). The objective is to determine the differences among the mean vectors for groups on the dependent variables, as well as differences among groups for the linear combinations of the dependent variables. This form of MANOVA extends Hotelling's $T^2$ to more than two groups; it is known as the one-way MANOVA, and it can be thought of as the MANOVA analog of the one-way $F$ situation. Results of the MANOVA produce four multivariate test statistics: Pillai's trace, Wilks's lambda ($\Lambda$), Hotelling's trace, and Roy's largest root. Usually results will not differ for the first three tests when applied to a two-group study; however, for studies involving more than two groups, tests may yield different results. The Wilks's $\Lambda$ is the test statistic reported most often in publications. The value of Wilks's $\Lambda$ ranges from 0 to 1. A small value of $\Lambda$ indicates statistically significant differences among the groups or treatment effects. Wilks's $\Lambda$, the associated $F$ value, hypotheses and error degrees of freedom, and the $p$ value are usually reported. A significant $F$ value is one that is greater than the critical value of $F$ at predetermined degrees of freedom for a preset level of significance. As a general rule, tables for critical values of $F$ and accompanying degrees of freedom are published as appendixes in many research and statistics books.

### *Factorial MANOVA*

Another common variation of multivariate procedures is known as the factorial MANOVA. In this

design, the effects of multiple factors on multiple dependent variables are examined. For example, the effects of geographic location and level of education on job satisfaction and attitudes toward work may be investigated via a factorial MANOVA. Geographic location with four levels and level of education with two levels are the factors. Geographic location could be coded as 1 = south, 2 = west, 3 = north, and 4 = east; level of education could be coded as 1 = college graduate and 0 = not college graduate. The MANOVA procedure will produce the main effects for each of the factors, as well as the interaction between the factors. For this example, three new dependent variables will be created to maximize group differences: one dependent variable to maximize the differences in geographic location and the linear combination of job satisfaction and attitudes toward work; one dependent variable to maximize the differences in education and the linear combination of job satisfaction and attitudes toward work; and another dependent variable to maximize separation among the groups for the interaction between geographic location and level of education. As in the previous designs, the factorial MANOVA produces Pillai's trace, Wilks's $\Lambda$, Hotelling's trace, and Roy's largest root. The multiple levels in factorial designs may produce slightly different values for the test statistics, even though these differences do not usually affect statistical significance. Wilks's $\Lambda$, associated $F$ statistic, degrees of freedom, and the $p$ value are usually reported in publications.

### K *Group MANOVA*

MANOVA designs with three or more groups are known as $K$ group MANOVAs. Like other multivariate designs, the null hypothesis tests whether differences between the mean vectors of $K$ groups on the combination of dependent variables are due to chance. As with the factorial design, the $K$ group MANOVA produces the main effects for each factor, as well as the interactions between factors. The same statistical tests and reporting requirements apply to the $K$ group situation as to the factorial MANOVA.

### *Doubly Multivariate Designs*

The purpose of doubly multivariate studies is to test for statistically significant group differences over time across a set of response variables measured at each time while accounting for the correlation among responses. A design would be considered doubly multivariate when multiple conceptually dissimilar dependent variables are measured across multiple time periods, as in a repeated measures study. For example, a study to compare problem-solving strategies of intrinsically and extrinsically motivated learners in different test situations could involve two dependent measures (score on a mathematics test and score on a reading test) taken at three different times (before a unit of instruction on problem solving, immediately following the instruction, and 6 weeks after the instruction) for each participant. Type of learner and test situation would be between-subjects factors and time would be a within-subjects factor.

### *Multivariate Analysis of Covariance*

A blend of analysis of covariance and MANOVA, called multivariate analysis of covariance (MANCOVA) allows the researcher to control for the effects of one or more covariates. MANCOVA allows the researcher to control for sources of variation within multiple variables. In the earlier example on attitudes toward the environment, the effects of concomitant variables such as number of people living in a household, age of head of household, gender, annual income, and education level can be statistically removed from the analysis with MANCOVA.

### *Factor Analysis*

MANOVA is useful as a data reduction procedure to condense a large number of variables into a smaller, more definitive set of hypothetical constructs. This procedure is known as *factor analysis*. Factor analysis is especially useful in survey research to reduce a large number of variables (survey items) to a smaller number of hypothetical variables by identifying variables that group or cluster together. For example, two or more dependent variables in a data set may measure the same entity or construct. If this is the case, the variables may be combined to form a new hypothetical variable. For example, a survey of students' attitudes toward work may include 40 related items, whereas a factor

analysis may reveal three underlying hypothetical constructs.

### *Discriminant Analysis*

A common use of discriminant analysis (DA) is prediction of group membership by maximizing the linear combination of multiple quantitative independent variables that best portrays differences among groups. For example, a college may wish to group incoming students based on their likelihood of being graduated. This is called predictive DA. Also, DA is used to describe differences among groups by identifying *discriminant functions* based on uncorrelated linear combinations of the independent variables. This technique may be useful following a MANOVA analysis and is called descriptive DA.

*Marie Kraska*

***See also*** Analysis of Variance (ANOVA); Discriminant Analysis; Multivariate Normal Distribution; Principal Components Analysis; Random Sampling; Repeated Measures Design

### Further Readings

Green, S. B., & Salkind, N. J. (2008). *Using SPSS for Windows and Macintosh: Analyzing and understanding data* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Johnson, R. A., & Wichern, D. W. (2002). *Applied multivariate statistical analysis* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Rencher, A. C. (2002). *Methods of multivariate analysis* (2nd ed.). San Francisco: Wiley.

Stevens, J. P. (2001). *Applied multivariate statistics for the social sciences* (4th ed.). Hillsdale, NJ: Lawrence Erlbaum.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). Boston: Allyn and Bacon.

Tatsuoka, M. M. (1988). *Multivariate analysis: Techniques for educational and psychological research* (2nd ed.). New York: Macmillan.

Timm, N. H. (2002). *Applied multivariate analysis*. New York: Springer-Verlag.

Weerahandi, S. (2004). *Generalized inference in repeated measures: Exact methods in MANOVA and mixed models*. San Francisco: Wiley.

## Multivariate Normal Distribution

One of the most familiar distributions in statistics is the normal or Gaussian distribution. It has two parameters, corresponding to the first two moments (mean and variance). Once these parameters are known, the distribution is completely specified. The multivariate normal distribution is a generalization of the normal distribution and also has a prominent role in probability theory and statistics. Its parameters include not only the means and variances of the individual variables in a multivariate set but also the correlations between those variables. The success of the multivariate normal distribution is due to its mathematical tractability and to the *multivariate central limit theorem*, which states that the sampling distributions of many multivariate statistics are normal, regardless of the parent distribution. Thus, the multivariate normal distribution is very useful in many statistical problems, such as multiple linear regressions and sampling distributions.

### Probability Density Function

If $\mathbf{X} = (X_1, \ldots, X_n)'$ is a multivariate normal random vector, denoted $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$ or $\mathbf{X} \sim N_n(\boldsymbol{\mu}, \Sigma)$, then its density is given by

$$f_{\mathrm{X}}(\mathbf{x}) = \frac{1}{|\Sigma|^{1/2}(2\pi)^{n/2}} \exp\left\{-\frac{1}{2}\left((\mathbf{x}-\boldsymbol{\mu})'\overset{-1}{\Sigma}(\mathbf{x}-\boldsymbol{\mu})\right)\right\},$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)' = E(\mathbf{X})$ is a vector whose components are the expectations $E(X_1), \ldots, E(X_n)$ and $\Sigma$ is the nonsingular variance–covariance matrix $(n \times n)$ whose diagonal terms are variances and off-diagonal terms are covariances:

$$\Sigma = V(\mathbf{X}) = E\left[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\{\boldsymbol{\mu})'\right] = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \ldots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \ldots & \sigma_n^2 \end{bmatrix}.$$

Note that the covariance matrix $\Sigma$ is symmetric and positive definite. The $(i,j)$th element is given by $\sigma_{ij} = E[(X_i - \mu_i)(X_j - \mu_j)]$ and $\sigma_i^2 \equiv \sigma_{ii} = V(X_i)$.

An important special case of the multivariate normal distribution is the bivariate normal. If $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2(\boldsymbol{\mu}, \Sigma)$, where $\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$, $\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$ and $\rho \equiv Corr(X_1, X_2) = \frac{\sigma_{12}}{\sigma_1\sigma_2}$, then the bivariate density is given by

$$f_{X_1,X_2}(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{(1-\rho^2)}} \exp\left\{ -\frac{\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2}{2(1-\rho^2)} \right\}.$$

Let $\mathbf{X} = (X_1, X_2)'$; the joint density can be rewritten in matrix notation as

$$f_X(\mathbf{x}) = \frac{1}{2\pi|\Sigma|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2}\left( (\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) \right) \right\}.$$

## Multivariate Normal Density Contours

The contour levels of $f_X(\mathrm{x})$, that is, the set of points in $\mathbb{R}^n$ for which $f_X(\mathrm{x})$ is constant, satisfy

$$(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) = c^2.$$

These surfaces are $n$-dimensional ellipsoids centered at $\boldsymbol{\mu}$, whose axes of symmetry are given by the principal components (the eigenvectors) of $\Sigma$. Specifically, the length of the ellipsoid along the $i$th axis is $c\sqrt{\lambda_i}$, where $\lambda_i$ is the $i$th eigenvalue associated with the eigenvector $e_i$ (recall that eigenvectors $e_i$ and eigenvalues $\lambda_i$ are solutions to $\Sigma e_i = \lambda_i e_i$ for $i = 1, \ldots, n$).

## Some Basic Properties

The following list presents some important properties involving the multivariate normal distribution.

*1.* The first two moments of a multivariate normal distribution, namely $\boldsymbol{\mu}$ and $\Sigma$, completely characterize the distribution. In other words, if $\mathbf{X}$ and $\mathbf{Y}$ are both multivariate normal with the same first two moments, then they are similarly distributed.

*2.* Let $\mathbf{X} = (X_1, \ldots, X_n)'$ be a multivariate normal random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\Sigma$, and let $\boldsymbol{\alpha}' = (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n/\mathbf{0}$. The linear combination $Y = \boldsymbol{\alpha}'\mathbf{X} = \alpha_1 X_1 + \cdots + \alpha_n X_n$ is normal with mean $E(Y) = \boldsymbol{\alpha}'\boldsymbol{\mu}$ and variance $V(Y) = \boldsymbol{\alpha}'\Sigma\boldsymbol{\alpha} = \sum_{i=1}^{n} \alpha_i^2 V(X_i) + \sum_{i \neq j} \sum \alpha_i \alpha_j Cov(X_i, X_j)$. Also, if $\boldsymbol{\alpha}'\mathbf{X}$ is normal with mean $\boldsymbol{\alpha}'\boldsymbol{\mu}$ and variance $\boldsymbol{\alpha}'\Sigma\boldsymbol{\alpha}$ for all possible $\boldsymbol{\alpha}$, then $\mathbf{X}$ must be a multivariate normal random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\Sigma$ ($\mathbf{X} \sim N_n(\boldsymbol{\mu}, \Sigma)$).

*3.* More generally, let $\mathbf{X} = (X_1, \ldots, X_n)'$ be a multivariate normal random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\Sigma$, and let $\mathrm{A} \in \mathbb{R}^{m \times n}$ be a full rank matrix with $m \leq n$, the set of linear combinations $\mathbf{Y} = (Y_1, \ldots, Y_m)' = \mathbf{AX}$ is multivariate normally distributed with mean $\mathbf{A}\boldsymbol{\mu}$ and covariance matrix $\mathbf{A}\Sigma\mathbf{A}'$. Also, if $\mathbf{Y} = \mathbf{AX} + \mathbf{b}$ where $\mathbf{b}$ is a $m \times 1$ vector of constants, then $\mathbf{Y}$ is multivariate normally distributed with mean $\mathbf{A}\boldsymbol{\mu} + \mathbf{b}$ and covariance matrix $\mathbf{A}\Sigma\mathbf{A}'$.

*4.* If $X_i$ and $Y_j$ are jointly normally distributed, then they are independent if and only if $Cov(Y_i, Y_j) = 0$. Note that it is not necessarily true that uncorrelated univariate normal random variables are independent. Indeed, two random variables that are marginally normally distributed may fail to be jointly normally distributed.

*5.* Let $\mathbf{Z} = (Z_1, \ldots, Z_n)'$ where $Z_i \sim^{i.i.d.} N(0,1)$ (where i.i.d. = independent and identically distributed). $\mathbf{Z}$ is said to be standard multivariate normal, denoted $\mathbf{Z} \sim N(\mathbf{0}, \mathbf{I}_n)$, and it can be shown that $E[\mathbf{Z}] = 0$ and $V(\mathbf{Z}) = \mathbf{I}_n$, where $\mathbf{I}_n$ denotes the unit matrix of order $n$. The joint density of vector $\mathbf{Z}$ is given by

$$f_Z(\mathbf{z}) = \prod_{i=1}^{n} f_{Z_i}(z_i) = (2\pi)^{-n/2} \exp\left\{ -\frac{1}{2}\mathbf{z}'\mathbf{z} \right\}.$$

The density $f_Z(\mathbf{z})$ is symmetric and unimodal with mode equal to zero. The contour levels of $f_Z(\mathbf{z})$, that is, the set of points in $\mathbb{R}^n$ for which $f_Z(\mathbf{z})$ is constant, are defined by

$$\mathbf{z}'\mathbf{z} = \sum_{i=1}^{n} z_i^2 = c^2,$$

where $c \geq 0$. The contour levels of $f_Z(z)$ are concentric circles in $\mathbb{R}^n$ centered at zero.

6. If $Y_1, \ldots, Y_n \sim^{ind} N(\mu_i, \sigma_i^2)$, then $\sigma_{ij} = 0$ for all $i \neq j$, and it follows that $\Sigma$ is a diagonal matrix. Thus, if

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix},$$

then

$$\overset{-1}{\Sigma} = \begin{pmatrix} 1/\sigma_1^2 & 0 & \cdots & 0 \\ 0 & 1/\sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/\sigma_n^2 \end{pmatrix},$$

so that

$$(\mathbf{y} - \boldsymbol{\mu})' \overset{-1}{\Sigma} (\mathbf{y} - \boldsymbol{\mu}) = \sum_{j=1}^{n} (y_j - \mu_j)^2 \Big/ \sigma_j^2.$$

Note also that, as $\Sigma$ is diagonal, we have

$$|\Sigma| = \sigma_1^2 \sigma_2^2 \ldots \sigma_n^2.$$

The joint density becomes

$$\begin{aligned} f_Y(y) &= \prod_{i=1}^{n} f_{Y_i}(y_i; \mu_i, \sigma_i^2) \\ &= \prod_{i=1}^{n} \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left\{ -\frac{1}{2} \left( \frac{y_i - \mu_i}{\sigma_i} \right)^2 \right\}. \end{aligned}$$

Thus, $f_Y(y)$ reduces to the product of univariate normal densities.

## Moment Generating Function

Let $\mathbf{Z} = (Z_1, \ldots, Z_n)'$ where $Z_i \sim^{i.i.d.} N(0, 1)$. As previously seen, $\mathbf{Z} \sim N(\mathbf{0}, \mathbf{I}_n)$ is referred to as a standard multivariate normal vector, and the density of $\mathbf{Z}$ is given by

$$f_Z(z) = (2\pi)^{-n/2} \exp\left\{ -\frac{1}{2} \mathbf{z}'\mathbf{z} \right\}.$$

The moment generating function of $\mathbf{Z}$ is obtained as follows:

$$\begin{aligned} M_Z(t) &= E\left[ e^{t'Z} \right] = (2\pi)^{-n/2} \int_n \mathbb{R}^n \exp\{\mathbf{t}'\mathbf{z} - \mathbf{z}'\mathbf{z}/2\} d\mathbf{z} \\ &= \prod_{i=1}^{n} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{ t_i z_i - \frac{1}{2} z_i^2 \right\} dz_i \\ &= \prod_{i=1}^{n} M_{Z_i}(t_i) \\ &= E\left[ e^{t_1 Z_1} \right] E\left[ e^{t_2 Z_2} \right] \ldots E\left[ e^{t_n Z_n} \right] \\ &= \exp\left\{ \frac{1}{2} t_1^2 \right\} \exp\left\{ \frac{1}{2} t_2^2 \right\} \ldots \exp\left\{ \frac{1}{2} t_n^2 \right\} \\ &= \exp\left\{ \frac{1}{2} \sum_{i=1}^{n} t_i^2 \right\} \\ &= \exp\left\{ \frac{1}{2} \mathbf{t}'\mathbf{t} \right\} \end{aligned}.$$

To obtain the moment generating function of the generalized location-scale family, let $\mathbf{X} = \{\boldsymbol{\mu} + \Sigma^{1/2}\mathbf{Z}$ where $\Sigma^{1/2}\Sigma^{1/2} = \Sigma$ ($\Sigma^{1/2}$ is obtained via the Cholesky decomposition of $\Sigma$) so that $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$. Hence,

$$\begin{aligned} M_X(t) &= E\left[ e^{t'X} \right] \\ &= E\left[ \exp\left\{ \mathbf{t}'\boldsymbol{\mu} + \mathbf{t}' \overset{1/2}{\Sigma} \mathbf{Z} \right\} \right] \\ &= e^{t'\mu} E\left[ \exp\left\{ \mathbf{t}' \overset{1/2}{\Sigma} \mathbf{Z} \right\} \right] \\ &= e^{t'\mu} M_Z\left( \left( \overset{1/2}{\Sigma} \right)' \mathbf{t} \right) \\ &= e^{t'\mu} \exp\left\{ \frac{1}{2} \left( \left( \overset{1/2}{\Sigma} \right)' \mathbf{t} \right)' \left( \left( \overset{1/2}{\Sigma} \right)' \mathbf{t} \right) \right\} \\ &= \exp\left\{ \mathbf{t}'\boldsymbol{\mu} + \frac{1}{2} \mathbf{t}' \Sigma \mathbf{t} \right\} \end{aligned}.$$

## Simulation

To generate a sample of observations from a random variable $\mathbf{Z} \sim N(\mathbf{0}, \mathbf{I}_n)$, one should note that

each of the $n$ components of vector $\mathbf{Z}$ is independent and identically distributed standard univariate normal, for which simulation methods are well known. Let $\mathbf{X} = \boldsymbol{\mu} + \Sigma^{1/2}\mathbf{Z}$ where $\Sigma^{1/2}\Sigma^{1/2} =$ so that $\mathbf{X} \sim N(\{\boldsymbol{\mu}, \Sigma)$. Realizations of $\mathbf{X}$ can be obtained from the generated samples $\mathbf{z}$ as $\{\boldsymbol{\mu} + \Sigma^{1/2}\mathbf{z}$ where $\Sigma^{1/2}$ can be computed via the Cholesky decomposition.

## Cumulative Distribution Function

The cumulative distribution function is the probability that all values in the random vector $\mathbf{X}$ are less than or equal to the values in the random vector $\mathbf{x}(\Pr(\mathbf{X} \le \mathbf{x}))$. Although there is no closed form for the cumulative distribution function of the multivariate normal, it can be calculated numerically by generating a large sample of observations and computing the fraction that satisfies $\mathbf{X} \le \mathbf{x}$.

## Marginal Distributions

Let $\mathbf{Y} = (Y_1, \ldots, Y_n)' \sim N(\boldsymbol{\mu}, \Sigma)$. Suppose that $\mathbf{Y}$ is partitioned into two subvectors, $\mathbf{Y}_{(1)}$ and $\mathbf{Y}_{(2)}$, so that we can write $\mathbf{Y} = \left(\mathbf{Y}'_{(1)}, \mathbf{Y}'_{(2)}\right)'$, where $\mathbf{Y}_{(1)} = (Y_1, \ldots, Y_P)'$ and $\mathbf{Y}_{(2)} = \left(Y_{p+1}, \ldots, Y_n\right)'$. Let $\mu$ and $\Sigma$ be independent and identically distributed partitioned accordingly, that is, $\boldsymbol{\mu} = \left(\boldsymbol{\mu}'_{(1)}, \boldsymbol{\mu}'_{(2)}\right)'$ and

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \quad \Sigma_{21} = \Sigma'_{12},$$

where $\boldsymbol{\mu}_{(i)} = E\left[\mathbf{Y}_{(i)}\right]$, $\Sigma_{ii} = V\left[\mathbf{Y}_{(i)}\right]$, $i = 1, 2$, and $\Sigma_{12} = Cov\left(\mathbf{Y}_{(1)}, \mathbf{Y}_{(2)}\right)$.

Then, it can be shown that the distributions of the two subvectors $\mathbf{Y}_{(1)}$ and $\mathbf{Y}_{(2)}$ are multivariate normal, defined as follows:

$$\mathbf{Y}_{(1)} \sim N(\boldsymbol{\mu}_1, \Sigma_{11}) \text{ and } \mathbf{Y}_{(2)} \sim N(\boldsymbol{\mu}_2, \Sigma_{22}).$$

This result means that

- Each of the $Y_i$s is univariate normal.
- All possible subvectors are multivariate normal.
- All marginal distributions are multivariate normal.

Moreover, if $\Sigma_{12} = 0$ ($\mathbf{Y}_{(1)}$ and $\mathbf{Y}_{(2)}$ are uncorrelated), then $\mathbf{Y}_{(1)}$ and $\mathbf{Y}_{(2)}$ are statistically independent. Recall that the covariance of two independent random variables is always zero but the opposite need not be true. Thus, $\mathbf{Y}_{(1)}$ and $\mathbf{Y}_{(2)}$ are statistically independent if and only if $\Sigma_{12} = \Sigma'_{12} = 0$.

## Conditional Normal Distributions

Let $\mathbf{Y} = (Y_1, \ldots, Y_n)' \sim N(\boldsymbol{\mu}, \Sigma)$. Suppose that $\mathbf{Y}$ is partitioned into two subvectors, $\mathbf{Y}_{(1)}$ and $\mathbf{Y}_{(2)}$, in the same manner as in the previous section.

The conditional distribution of $\mathbf{Y}_{(1)}$ given $\mathbf{Y}_{(2)}$ is multivariate normal characterized by its mean and covariance matrix as follows:

$$\left(\mathbf{Y}_{(1)} \middle| \mathbf{Y}_{(2)} = \mathbf{y}_{(2)}\right) \Sigma$$
$$N\left(\boldsymbol{\mu}_{(1)} + \Sigma_{12}\Sigma_{22}^{-1}\left(\mathbf{y}_{(2)} - \boldsymbol{\mu}_{(2)}\right), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right).$$

To verify this, let $\mathbf{X} = \mathbf{Y}_{(1)} - \mu_{(1)} - \Sigma_{12}\Sigma_{22}^{-1}\left(\mathbf{Y}_{(2)} - \boldsymbol{\mu}_{(2)}\right)$. As a linear combination of the elements of $\mathbf{Y}$, $\mathbf{X}$ is also normally distributed with mean zero and variance $V(\mathbf{X}) = E[\mathbf{X}\mathbf{X}']$, given by:

$$\begin{aligned} &E[\mathbf{X}\mathbf{X}'] = \\ &E\left[\left\{\left(\mathbf{Y}_{(1)} - \mu_{(1)}\right) - \Sigma_{12}\Sigma_{22}^{-1}\left(\mathbf{Y}_{(2)} - \boldsymbol{\mu}_{(2)}\right)\right\}\right. \\ &\quad \left.\times\left\{\left(\mathbf{Y}_{(1)} - \boldsymbol{\mu}_{(1)}\right)' - \left(\mathbf{Y}_{(2)} - \boldsymbol{\mu}_{(2)}\right)'\Sigma_{22}^{-1}\Sigma_{12}\right\}\right] \\ &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \\ &\quad + \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{22}\Sigma_{22}^{-1}\Sigma_{21} \\ &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \end{aligned}$$

Moreover, if we consider $\left(\mathbf{X}', \mathbf{Y}'_{(2)}\right)'$, which is also multivariate normal, we obtain the following covariance term:

$$\begin{aligned} &E[(\mathbf{X} - 0)\left(\mathbf{Y}_{(2)} - \boldsymbol{\mu}_{(2)}\right)'] \\ &= E[\left(\mathbf{Y}_{(1)} - \boldsymbol{\mu}_{(1)} - \Sigma_{12}\Sigma_{22}^{-1}\left(\mathbf{Y}_{(2)} - \boldsymbol{\mu}_{(2)}\right)\right) \\ &\quad \left(\mathbf{Y}_{(2)} - \boldsymbol{\mu}_{(2)}\right)'] = \Sigma_{12} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{22} = 0. \end{aligned}$$

This implies that $\mathbf{X}$ and $\mathbf{Y}_{(2)}$ are independent, and we can write

$$\begin{bmatrix}\mathbf{X}\\ \mathbf{Y}_{(2)}\end{bmatrix}\sim N\left(\begin{bmatrix}\mathbf{0}\\ \boldsymbol{\mu}_{(2)}\end{bmatrix},\begin{bmatrix}\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & \mathbf{0}\\ \mathbf{0} & \Sigma_{22}\end{bmatrix}\right).$$

As

$$\mathbf{Y}_{(1)}=\mathbf{X}+\boldsymbol{\mu}_{(1)}+\Sigma_{12}\Sigma_{22}^{-1}\left(\mathbf{Y}_{(2)}-\boldsymbol{\mu}_{(2)}\right),$$

conditional on $\mathbf{Y}_{(2)}=\mathbf{y}_{(2)}$, $\mathbf{Y}_{(1)}$ is normally distributed with mean

$$\underbrace{E(\mathbf{X})}_{0}+\boldsymbol{\mu}_{(1)}+\Sigma_{12}\ \Sigma_{22}^{-1}\left(\mathbf{y}_{(2)}-\boldsymbol{\mu}_{(2)}\right)$$

and variance $V(\mathbf{X})=\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

In the bivariate normal case, we replace $\mathbf{Y}_{(1)}$ and $\mathbf{Y}_{(2)}$ by $Y_1$ and $Y_2$, and the conditional distribution of $Y_1$ given $Y_2$ is normal as follows:

$$Y_1/(Y_2=y_2)\sim N\left(\mu_1+\rho\frac{\sigma_1}{\sigma_2}(y_2-\mu_2),\sigma_1^2\left(1-\rho^2\right)\right).$$

Another special case is with $\mathbf{Y}_{(1)}=Y_1$ and $\mathbf{Y}_{(2)}=(Y_2,\ldots,Y_n)'$ so that

$$\left(Y_1\middle|\mathbf{Y}_{(2)}=\mathbf{y}_{(2)}\right)\sim N\left(\mu_1+\Sigma_{12}\Sigma_{22}^{-1}\left(\mathbf{y}_{(2)}-\boldsymbol{\mu}_{(2)}\right),\ \Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right).$$

The mean of this particular conditional distribution ($\mu_1+\Sigma_{12}\Sigma_{22}^{-1}\left(\mathbf{y}_{(2)}-\boldsymbol{\mu}_{(2)}\right)$) is referred to as a multiple regression function of $Y_1$ on $\mathbf{Y}_2$, with the regression coefficient vector being $\boldsymbol{\beta}=\Sigma_{12}\Sigma_{22}^{-1}$.

## Partial Correlation

Let $\mathbf{Y}=(Y_1,\ldots,Y_n)'\sim N(\boldsymbol{\mu},\Sigma)$. The covariance (correlation) between two of the univariate random variables in $\mathbf{Y}$, say $Y_i$ and $Y_j$, is determined from the $(i,j)$th entry of $\Sigma$. The partial correlation concept considers the correlation of $Y_i$ and $Y_j$ when conditioning on a set of other variables in $\mathbf{Y}$. Let $\mathbf{C}=\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$, which is the variance of the conditional distribution of $Y_{(1)}$ given $Y_{(2)}=y_{(2)}$, and denote the $(i,j)$th element of $\mathbf{C}$ by $\sigma_{ij|(p+1,\ldots,n)}$, where $p$ is the dimension of the vector $Y_{(1)}$.

The partial correlation of $Y_i$ and $Y_j$, given $\mathbf{Y}_{(2)}$, is defined by

$$\rho_{ij|(p+1,\ldots,n)}=\frac{\sigma_{ij|(p+1,\ldots,n)}}{\sqrt{\sigma_{ii|(p+1,\ldots,n)}\sigma_{jj|(p+1,\ldots,n)}}}.$$

## Testing for Multivariate Normality

If a vector is multivariate normally distributed, each individual component follows a univariate Gaussian distribution. However, univariate normality of each variable in a set is not sufficient for multivariate normality. Therefore, to test the adequacy of the multivariate normality assumption, several authors have presented a battery of tests consistent with the multivariate framework. For this purpose, Kanti V. Mardia proposed multivariate extensions of the skewness and kurtosis statistics, which are extensively used in the univariate framework to test for normality.

Let $\{\mathbf{X}_i\}_{i=1}^m=\{X_{1i},\ldots,X_{ni}\}_{i=1}^m$ be the $i$th random sample from an $n$-variate distribution, $\overline{\mathbf{X}}$ the vector of sample means, and $\mathbf{S}_X$ the sample covariance matrix. The $n$-variate skewness and kurtosis statistics denoted respectively by $b_{1,n}$ and $b_{2,n}$ are defined as follows:

$$b_{1,n}=\frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\left(\left(\mathbf{X}_i-\overline{\mathbf{X}}\right)'\mathbf{S}_X^{-1}\left(\mathbf{X}_j-\overline{\mathbf{X}}\right)\right)^3$$

$$b_{2,n}=\frac{1}{m}\sum_{i=1}^{m}\left(\left(\mathbf{X}_i-\overline{\mathbf{X}}\right)'\mathbf{S}_X^{-1}\left(\mathbf{X}_i-\overline{\mathbf{X}}\right)\right)^2.$$

We can verify that these statistics reduce to the well-known univariate skewness and kurtosis for $n=1$.

Based of their asymptotic distributions, the multivariate skewness and kurtosis defined above can be used to test for multivariate normality. If $\mathbf{X}$ is multivariate normal, then $b_{1,n}$ and $b_{2,n}$ have expected values 0 and $n(n+2)$. It can be also shown that for large $m$, the limiting distribution of $(m/6)b_{1,n}$ is a chi square with $n(n+1)(n+2)/6$ degrees of freedom, and the limiting distribution of $\sqrt{m}\left(b_{2,n}-n(n+2)\right)/\sqrt{8n(n+2)}$ is $N(0,1)$.

## The Gaussian Copula Function

Recently in multivariate modeling, much attention has been paid to copula functions. A copula is a function that links an $n$-dimensional distribution function to its one-dimensional margins and is itself a continuous distribution function characterizing the dependence structure of the model. Sklar's theorem states that under appropriate conditions, the joint density can be written as a product of the marginal densities and the copula density. Several copula families are available that can incorporate the relationships between random variables. Among these families, the Gaussian copula encodes dependence in precisely the same way as the multivariate normal distribution does, using only pairwise correlations among the variables. However, it does so for variables with arbitrary margins. A multivariate normal distribution arises whenever univariate normal margins are linked through a Gaussian copula. The Gaussian copula function is defined by

$$C(u_1, u_2, \ldots, u_n) = \Phi_\rho^n\left(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \ldots, \Phi^{-1}(u_n)\right),$$

where $\Phi_\rho^n$ denotes the joint distribution function of the $n$-variate standard normal distribution with linear correlation matrix $\rho$, and $\Phi^{-1}$ denotes the inverse of the univariate standard normal distribution function. In the bivariate case, the copula expression can be written as

$$C(u, v) = \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2}\frac{x^2 - 2\rho xy + y^2}{(1-\rho^2)}\right\} dxdy,$$

where $\rho$ is the usual linear correlation coefficient of the corresponding bivariate normal distribution.

## Multiple Linear Regression and Sampling Distribution

In this section, the multiple regression model and the associated sampling distribution are presented as an illustration of the usefulness of the multivariate normal distribution in statistics.

Consider the following multiple linear regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_q X_q + \varepsilon,$$

where $Y$ is the response variable, $(X_1, X_2, \ldots, X_q)'$ is a vector representing a set of $q$ explanatory variables, and $\varepsilon$ is the error term. Note that the simple linear regression model is a special case with $q = 1$.

Suppose that we have $n$ observations on $Y$ and on each of the explanatory variables, that is,

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \cdots + \beta_q X_{1q} + \varepsilon_1 \\ Y_2 &= \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + \cdots + \beta_q X_{2q} + \varepsilon_2 \\ &\vdots \quad \vdots \\ Y_n &= \beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2} + \cdots + \beta_q X_{nq} + \varepsilon_n \end{aligned},$$

where $E(\varepsilon_i) = 0$, $\mathrm{var}(\varepsilon_j) = \sigma^2$ and $\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$.

We can rewrite

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1q} \\ 1 & X_{21} & X_{22} & \cdots & X_{2q} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nq} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_q \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

or else

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\mathrm{cov}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_n$.

Note that in the previous expression, $\boldsymbol{\varepsilon}$ is a multivariate normal random vector whereas $\mathbf{X}\boldsymbol{\beta}$ is a vector of constants. Thus, $\mathbf{Y}$ is a linear combination of a multivariate normally distributed vector. It follows that $\mathbf{Y}$ is also multivariate normal with mean $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ and covariance $\mathrm{cov}(\mathbf{Y}) = \sigma^2 \mathbf{I}_n$.

The goal of an analysis of data of this form is to estimate the regression parameter $\boldsymbol{\beta}$. The least

squares estimate of $\boldsymbol{\beta}$ is found by minimizing the sum of squared deviations

$$\sum_{j=1}^{n} \left(Y_j - \beta_0 - \beta_1 X_{j1} + \beta_2 X_{j2} + \cdots + \beta_q X_{jq}\right)^2.$$

In matrix terms, this sum of squared deviations may be written

$$(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}.$$

If $\mathbf{X}$ is of full rank, then the least squares estimator for $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

It is of interest to characterize the probability distribution of an estimator. If we refer to the multiple regression problem, the estimator $\hat{\boldsymbol{\beta}}$ depends on the response $\mathbf{Y}$. Therefore, the properties of its distribution will depend on those of $\mathbf{Y}$. More specifically the distribution of $\hat{\boldsymbol{\beta}}$ is multivariate normal as follows:

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}).$$

This result can be used to obtain estimated standard errors for the components of $\hat{\boldsymbol{\beta}}$, that is, estimates of the standard deviation of the sampling distributions of each component of $\hat{\boldsymbol{\beta}}$.

*Chiraz Labidi*

***See also*** Central Limit Theorem; Coefficients of Correlation, Alienation, and Determination; Copula Functions; Kurtosis; Multiple Regression; Normal Distribution; Normality Assumption; Partial Correlation; Sampling Distributions

**Further Readings**

Balakrishnan, N., & Nevrozov, V. B. (2003). *A primer on statistical distributions*. New York: Wiley.

Evans, M., Hastings, N., & Peacock, B. (2000). *Statistical distributions*. New York: Wiley.

Kotz, S., Balakrishnan, N., & Johnson, N. L. (2000). *Continuous multivariate distributions, Volume 1: Models and applications*. New York: Wiley.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika, 57,* 519–530.

Paolella, M. S. (2007). *Intermediate probability: A computational approach*. New York: Wiley.

Rose, C., & Smith, M. D. (1996). The multivariate normal distribution. *Mathematica, 6,* 32–37.

Sklar, A. (1959). *N*-dimensional distribution functions and their margins. *Publications de l'Institut de Statistique de l'Université de Paris, 8,* 229–231.

Tong, Y. L. (1990). *The multivariate normal distribution*. New York: Springer-Verlag.

# N

## Narrative Research

Narrative research aims to explore and conceptualize human experience as it is represented in textual form. Aiming for an in-depth exploration of the meanings people assign to their experiences, narrative researchers work with small samples of participants to obtain rich and free-ranging discourse. The emphasis is on storied experience. Generally, this takes the form of interviewing people around the topic of interest, but it might also involve the analysis of written documents. Narrative research as a mode of inquiry is used by researchers from a wide variety of disciplines, which include anthropology, communication studies, cultural studies, economics, education, history, linguistics, medicine, nursing, psychology, social work, and sociology. It encompasses a range of research approaches including ethnography, phenomenology, grounded theory, narratology, action research, and literary analysis, as well as such interpretive stances as feminism, social constuctionism, symbolic interactionism, and psychoanalysis. This entry discusses several aspects of narrative research, including its epistemological grounding, procedures, analysis, products, and advantages and disadvantages.

### Epistemological Grounding

The epistemological grounding for narrative research is on a continuum of postmodern philosophical ideas in that there is a respect for the relativity and multiplicity of truth in regard to the human sciences. Narrative researchers rely on the epistemological arguments of such philosophers as Paul Ricoeur, Martin Heidegger, Edmund Husserl, Wilhelm Dilthey, Ludwig Wittgenstein, Mikhail Bakhtin, Jean-Francois Lyotard, and Hans-Georg Gadamer. Although narrative researchers differ in their view of the possibility of objectively conceived "reality," most agree with Donald Spence's distinction between narrative and historical truth. Factuality is of less interest than how events are understood and organized, and all knowledge is presumed to be socially constructed.

Ricoeur, in his seminal work *Time and Narrative,* argues that time is organized and experienced narratively; narratives bring order and meaning to the constantly changing flux. In its simplest form, our experience is internally ordered as "this happened, then that happened" with some (often causal) connecting link in between. Narrative is also central to how we conceive of ourselves; we create stories of ourselves to connect our actions, mark our identity, and distinguish ourselves from others.

Questions about *how* people construct themselves and others in various contexts, under various conditions, are the focus of narrative research. Narrative research paradigms, in contrast to hypothesis-testing ones, have as their aims describing and understanding rather than measuring and predicting, focusing on meaning rather than causation and frequency, interpretation rather than statistical analysis, and recognizing the importance of

language and discourse rather than reduction to numerical representation. These approaches are holistic rather than atomistic, concern themselves with particularity rather than universals, are interested in the cultural context rather than trying to be context-free, and give overarching significance to subjectivity rather than questing for some kind of objectivity.

Narrative research orients itself toward understanding human complexity, especially in those cases where the many variables that contribute to human life cannot be controlled. Narrative research aims to take into account—and interpretively account for—the multiple perspectives of *both* the researched and researcher.

Jerome Bruner has most championed the legitimization of what he calls "narrative modes of knowing," which privileges the particulars of lived experience rather than constructs about variables and classes. It aims for the understanding of lives in context rather than through a prefigured and narrowing lens. Meaning is not inherent in an act or experience but is constructed through social discourse. Meaning is generated by the linkages the participant makes between aspects of the life he or she is living and by the explicit linkages the researcher makes between this understanding and interpretation, which is meaning constructed at another level of analysis.

## Life Is a Story

One major presupposition of narrative research is that humans experience their lives in emplotted forms resembling stories or at least communicate about their experiences in this way. People use narrative as a form of constructing their views of the world; time itself is constructed narratively. Important events are represented as taking place through time, having roots in the past, and extending in their implications into the future. Life narratives are also contextual in that persons act within situational contexts that are both immediate and more broadly societal. The focus of research is on what the individuals think they are doing and why they think they are doing so. Behavior, then, is always understood in the individual's context, however he or she might construct it. Thus, narratives can be examined for personal meanings, cultural meanings, and the interaction between these.

Narrative researchers invite participants to describe in detail—tell the story of—either a particular event or a significant aspect or time of life (e.g., a turning point), or they ask participants to narrate an entire life story. Narration of experience, whether of specific events or entire life histories, involves the subjectivity of the actor, with attendant wishes, conflicts, goals, opinions, emotions, worldviews, and morals, all of which are open to the gaze of the researcher. Such narratives also either implicitly or explicitly involve settings that include those others who are directly involved in the events being related and also involve all those relationships that have influenced the narrator in ever-widening social circles. The person is assumed to be speaking from a specific position in culture and in historical time. Some of this positionality is reflected in the use of language and concepts with which a person understands her or his life. Other aspects of context are made explicit as the researcher is mindful of the person's experience of herself or himself in terms of gender, race, culture, age, social class, sexual orientation, nationality, etc. Participants are viewed as unique individuals with particularity in terms of social location; a person is not viewed as representative of some universal and interchangeable, randomly selected "subject."

People, however, do not "have" stories of their lives; they create them for the circumstance in which the story will be told. No two interviewers will obtain exactly the same story from an individual interviewee. Therefore, a thoroughly reflexive analysis of the parameters and influences on the interview situation replaces concern with reliability.

A narrative can be defined as a story of a sequence of events. Narratives are organized so as to place meanings retrospectively on events, with events described in such a way as to express the meanings the narrator wishes to convey. Narrative is a way of understanding one's own (and others') action, of organizing personal experience, both internal and external, into a meaningful whole. This involves attributing agency to the characters in the narrative and inferring causal links between the events. In the classic formulation, a narrative is an account with three components: a beginning, a middle, and an end. William Labov depicts all narratives as having clauses that

orient the reader to the story, tell about the events, or evaluate the story—that is, instruct the listener or reader as to how the story is to be understood. The evaluation of events is of primary interest to the narrative researcher because this represents the ways in which the narrator constructs a meaning (or set of meanings) within the narrative. Such meanings, however, are not viewed to be either singular or static. Some narrative theorists have argued that the process of creating an autobiographical narrative is itself transforming of self because the self that is fashioned in the present extends into the future. Thus, narrative research is viewed to be investigating a self that is alive and evolving, a self that can shift meanings, rather than a fixed entity.

Narrative researchers might also consider the ways in which the act of narration is performative. Telling a story constructs a self and might be used to accomplish a social purpose such as defending the self or entertaining someone. Thus, the focus is not only on the content of what is communicated in the narrative but also on how the narrator constructs the story and the social locations from which the narrator speaks. Society and culture also enable and constrain certain kinds of stories; meaning making is always embedded in the concepts that are culturally available at a particular time, and these might be of interest in a narrative research project. Narrative researchers, then, attend to the myriad versions of self, reality, and experience that the storyteller produces through the telling.

## Procedures

Narrative research begins with a conceptual question derived from existing knowledge and a plan to explore this question through the narratives of people whose experience might illuminate the question. Most narrative research involves personal interviews, most often individual, but sometimes in groups. Some narrative researchers might (also) use personal documents such as journals, diaries, memoirs, or films as bases for their analyses. Narrative research uses whatever storied materials are available or can be produced from the kinds of people who might have personal knowledge and experiences to bring to bear on the research question.

In interview-based designs, which are the most widespread form of narrative research, participants who fit into the subgroup of interest are invited to be interviewed at length (generally 1–4 hours). Interviews are recorded and then transcribed. The narrative researcher creates "experience-near" questions related to the conceptual question that might be used to encourage participants to tell about their experiences. This might be a request for a full life story or it might be a question about a particular aspect of life experience such as life transitions, important relationships, or responses to disruptive life events.

Narrative research meticulously attends to the process of the interview that is organized in as unstructured a way as possible. The narrative researcher endeavors to orient the participant to the question of interest in the research and then intervene only to encourage the participant to continue the narration or to clarify what seems confusing to the researcher. Inviting stories, the interviewer asks the participant to detail his or her experiences in rich and specific narration. The interviewer takes an empathic stance toward the interviewees, trying to understand their experience of self and world from their point of view. Elliott Mishler, however, points out that no matter how much the interviewer or researcher attempts to put aside his or her own biases or associations to the interview content, the researcher has impact on what is told and this must be acknowledged and reflected on.

Because such interviews usually elicit highly personal material, confidentiality and respect for the interviewee must be assured. The ethics of the interview are carefully considered in advance, during the interview itself and in preparation of the research report.

Narrative research questions tend to focus on individual, developmental, and social processes that reflect how experience is constructed both internally and externally. Addressing questions that cannot be answered definitively, narrative research embraces multiple interpretations rather than aiming to develop a single truth. Rooted in a postmodern epistemology, narrative approaches to research respect the relativity of knowing—the meanings of the participant filtered through the mind of the researcher with all its assumptions and a priori meanings. Knowledge is presumed to be constructed rather than discovered and is assumed

to be localized, perspectival, and occurring within intersubjective relationships to both participants and readers. "Method" then becomes not a set of procedures and techniques but ways of thinking about inquiry, modes of exploring questions, and creative approaches to offering one's constructed findings to the scholarly community. All communication is through language that is understood to be always ambiguous and open to interpretation. Thus, the analytic framework of narrative research is in hermeneutics, which is the science of imprecise and always shifting meanings.

## Analysis

The analysis of narrative research texts is primarily aimed at inductively understanding the meanings of the participant and organizing them at some more conceptual level of understanding. This might involve a close reading of an individual's interview texts, which includes coding for particular themes or extracting significant passages for discussion in the report. The researcher looks inductively for patterns, and the kinds of patterns recognized might reflect the researcher's prior knowledge about the phenomena. The process of analysis is one of piecing together data, making the invisible apparent, deciding what is significant and what is insignificant, and linking seemingly unrelated facets of experience together. Analysis is a creative process of organizing data so the analytic scheme will emerge. Texts are read multiple times in what Friedrich Schleiermacher termed a "hermeneutic circle," a process in which the whole illuminates the parts that in turn offers a fuller and more complex picture of the whole, which then leads to a better understanding of the parts, and so on.

Narrative researchers focus first on the voices *within* each narrative, attending to the layering of voices (subject positions) and their interaction, as well as the continuities, ambiguities, and disjunctions expressed. The researcher pays attention to both the content of the narration ("the told") and the structure of the narration ("the telling"). Narrative analysts might also pay attention to what is unsaid or unsayable by looking at the structure of the narrative discourse and markers of omissions. After each participant's story is understood as well as possible, cross-case analysis might be performed to discover patterns across individual narrative interview texts or to explore what might create differences between people in their narrated experiences.

There are many approaches to analyses, with some researchers focusing on meanings through content and others searching through deconstructing the use and structure of language as another set of markers to meanings. In some cases, researchers aim to depict the layers of experience detailed in the narratives, preserving the point of view, or voice, of the interviewee. At other times, researchers might try to go beyond what is said and regard the narrated text as a form of disguise; this is especially true when what is sought are unconscious processes or culturally determined aspects of experience that are embedded rather than conscious.

The linguistic emphasis in some branches of narrative inquiry considers the ways in which language organizes both thought and experience. Other researchers recognize the shaping function of language but treat language as transparent as they focus more on the content of meanings that might be created out of life events.

The purpose of narrative research is to produce a deep understanding of dynamic processes. No effort is made to generalize about populations. Thus, statistics, which aims to represent populations and the distribution of variables within them, have little or no place in narrative research. Rather, knowledge is viewed to be localized in the analysis of the particular people studied and generalization about processes that might apply to other populations is left to the reader. That is, in a report about the challenges of immigration in a particu-lar population, the reader might find details of the interactive processes that might illuminate the struggles of another population in a different locale—or even people confronting other life transitions.

Narrative research avoids having a predetermined theory about the person that the interview or the life-story is expected to support. Although no one is entirely free of preconceived ideas and expectations, narrative researchers try to come to their narrators as listeners open to the surprising variation in their social world and private lives. Although narrative researchers try to be as knowledgeable as possible about the themes that they

are studying to be maximally sensitive to nuances of meaning, they are on guard again inflicting meaning in the service of their own ends.

## Products

Reports of narrative research privilege the words of the participants, in what Clifford Geertz calls "thick description," and present both some of the raw data of the text as well as the analysis. Offering as evidence the contextualized words of the narrator lends credence to the analysis suggested by the researcher. The language of the research report is often near to experience as lived rather than as obscured by scientific jargon. Even Sigmund Freud struggled with the problem of making the study of experience scientific, commenting in 1893 that the nature of the subject was responsible for his works reading more like short stories than customary scientific reports. The aim of a narrative research report is to offer interpretation in a form that is faithful to the phenomena. In place of form-neutral "objectivized" language, many narrative researchers concern themselves with the poetics of their reports and strive to embody the phenomena in the language they use to convey their meanings. Narrative researchers stay respectful of their participants and reflect on how they are representing "the other" in the published report.

Some recent narrative research has concerned such topics as how people experience immigration, illness, identity, divorce, recovery from addictions, belief systems, and many other aspects of human experience. Any life experiences that people can narrate or represent become fertile ground for narrative research questions. The unity of a life resides in a construction of its narrative, a form in which hopes, dreams, despairs, doubts, plans, and emotions are all phrased.

Although narrative research is generally concerned with individuals' experience, some narrative researchers also consider narratives that particular collectives (societies, groups, or organiza-tions) tell about themselves, their histories, their dominant mythologies, and their aspirations. Just as personal narratives create personal identity, group narratives serve to bond a community and distinguish it from other collectives.

A good narrative research report will detail a holistic overview of the phenomena under study, capturing data from the inside of the actors with a view to understanding and conceptualizing their meaning making in the contexts within which they live. Narrative researchers recognize that many interpretations of their observations are possible and they argue their interpretive framework through careful description of what they have observed.

Narrative researchers also recognize that they themselves are narrators as they present their organization and interpretation of their data. They endeavor to make their work as interpreters transparent, writing about their own interactions with their participants and their data and remaining mindful of their own social location and personal predilections. This reflexive view of researcher as narrator opens questions about the representation of the other and the nature of interpretive authority, and these are addressed rather than elided.

## Advantages and Disadvantages

A major appeal of narrative research is the opportunity to be exploratory and make discoveries free of the regimentation of prefabricated hypotheses, contrived variables, control groups, and statistics. Narrative research can be used to challenge conceptual hegemony in the social sciences or to extend the explanatory power of abstract theoretical ideas. Some of the most paradigm-defining conceptual revolutions in the study of human experience have come from narrative research—Sigmund Freud, Erik Erikson, and Carol Gilligan being the most prominent examples. New narrative researchers, however, struggle with the vaguely defined procedures on which this research depends and with the fact that interesting results cannot be guaranteed in advance. Narrative research is also labor intensive, particularly in the analysis phase, where text must be read and reread as insights and interpretations develop.

Narrative research is not generalizable to populations but rather highlights the particularities of experience. Many narrative researchers, however, endeavor to place the individual narratives they present in a broader frame, comparing and contrasting their conclusions with the work of others with related concerns. All people are like all other people, like some other people—and also are unique. Readers of narrative research are invited

explicitly to apply what is learned to contexts that are meaningful to them.

Narrative research opens possibilities for social change by giving voice to marginalized groups, representing unusual or traumatic experiences that are not conducive to control group designs, and by investigating the ways in which social life (and attendant oppression) is mediated through metanarratives. Readers, then, are challenged to understand their own or their society's stories in new ways at both experiential and theoretical levels.

*Ruthellen Josselson*

***See also*** Case-Only Design; Case Study; Naturalistic Observation; Observational Research; Observations

**Further Readings**

Andrews, M. (2007). *Shaping history: Narratives of political change*. Cambridge, UK: Cambridge University Press.

Bruner, J. (1990). *Acts of meaning*. Cambridge, MA: Harvard University Press.

Clandinnin, J. (Ed.). (2007). *The handbook of narrative inquiry*. Thousand Oaks, CA: Sage.

Freeman, M. (1993). *Rewriting the self: History, memory, narrative*. London: Routledge.

Hinchman, L. P., & Hinchman, S. K. (Eds). (1997). *Memory, identity, community: The idea of narrative in the human sciences*. Albany: State University of New York.

Josselson, R., Lieblich, A., & McAdams, D. P. (Eds.). (2003). *Up close and personal: The teaching and learning of narrative research*. Washington, DC: APA Books.

Lieblich, A., Tuval-Mashiach, R., & Zilber, T. (1998). *Narrative research: Reading, analysis and interpretation*. Thousand Oaks, CA: Sage.

McAdams, D. P. (1993). *The stories we live by: Personal myths and the making of the self*. New York: Morrow.

Mishler, E. (2004). Historians of the self: Restorying lives, revising identities. *Research in Human Development, 1*, 101–121.

Polkinghorne, D. (1988) *Narrative knowing and the human sciences*. Albany: State University of New York.

Rosenwald, G. C., & Ochberg, R. L. (Eds.). (1992). *Storied lives: The cultural politics of self-understanding*. New Haven, CT: Yale University Press.

Sarbin, T. R. (1986). *Narrative psychology: The storied nature of human conduct*. New York: Praeger.

Spence, D. (1982). *Narrative truth and historical truth*. New York: Norton.

# National Council on Measurement in Education

The National Council on Measurement in Education (NCME) is the sole professional association devoted to the scientific study and improvement of educational measurement. Originally founded in the United States in February 1938, it currently includes members from countries throughout the world. As of November 2008, it has approximately 2,000 members, including professional members and graduate students. Professional members work in university settings; national, state, and local government settings (typically departments of education); testing companies; and other industrial settings. Graduate student members are typically enrolled in graduate programs housed in schools of education or departments of psychology.

Since its founding, the organization's name has changed several times. Originally founded as the National Association of Teachers of Educational Measurement, its name was changed to National Council on Measurements Used in Education in 1943. The organization took its current name in 1960.

The mission statement of NCME affirms that it is "incorporated exclusively for scientific, educational, literary, and charitable purposes." In addition, the mission statement describes two major purposes of NCME. The first is the "encouragement of scholarly efforts to: Advance the science of measurement in the field of education; improve measurement instruments and procedures for their administration, scoring, interpretation, and use; and improve applications of measurement in assessment of individuals and evaluations of educational programs." The second focuses on dissemination: "Dissemination of knowledge about: Theory, techniques, and instrumentation available for measurement of educationally relevant human, institutional, and social characteristics; procedures appropriate to the interpretation and use of such techniques and instruments; and applications of educational measurement in individual and group evaluation studies."

These purpose statements underscore the NCME's focus on supporting research on educational tests and testing, development of improved

assessments in education, and disseminating information regarding new developments in educational testing and the proper use of tests. To accomplish these goals, NCME hosts an annual conference each year (jointly scheduled with the annual conference of the American Educational Research Association), publishes two highly regarded journals covering research and practice in educational measurement, and partners with other professional organizations to develop and disseminate guidelines and standards for appropriate educational assessment practices and to further the understanding of the strengths and limitations of educational tests. In the following sections, some of the most important activities of NCME are described.

## Dissemination Activities

The NCME publishes two peer-reviewed journals, both of which have four volumes per year. The first is the Journal of Educational Measurement (JEM), which was first published in 1963. JEM publishes original research related to educational assessment, particularly advances in statistical techniques such as equating tests (maintaining score scales over time), test calibration (e.g., using item response theory), and validity issues related to appropriate test development and use (e.g., techniques for evaluating item and test bias). It also publishes reviews of books related to educational measurement and theoretical articles related to major issues and developments in educational measurement (e.g., reliability and validity theory). The second journal published by NCME is Educational Measurement: Issues and Practice (EM:IP), which was first published in 1982. EM:IP focuses on more applied issues, typically less statistical in nature, that are of broad interest to measurement practitioners. According to the EM:IP page on the NCME website, the primary purpose of EM:IP is "to promote a better understanding of educational measurement and to encourage reasoned debate on current issues of practical importance to educators and the public. EM:IP also provides one means of communication among NCME members and between NCME."

In addition to the two journals, NCME also publishes the *Instructional Topics in Educational Measurement Series* (ITEMS), which are instructional units on specific measurement topics of interest to measurement researchers and practitioners. The ITEMS units, which first appeared in 1987, are available for free and can be downloaded from the ITEMS page on the NCME website. As of 2010, there are 22 modules covering a broad range of topics such as how to equate tests, evaluate differential item functioning, or set achievement level standards on tests.

NCME has also partnered with other organizations to publish books and other materials designed to promote fair or improved practices related to educational measurement. The most significant partnership has been with the Joint Committee on Testing Standards, which produced the *Standards for Educational and Psychological Testing* in 1999 as well as the four previous versions of those standards (in 1954, 1966, 1974, and 1985). NCME also partnered with the American Council on Education to produce four versions of the highly acclaimed book *Educational Measurement*. Two books on evaluating teachers were also sponsored by NCME: the *Handbook of Teacher Evaluation* and the *New Handbook of Teacher Evaluation: Assessing Elementary and Secondary School Teachers.*

In addition to publishing journals, instructional modules, and books, NCME has also partnered with other professional organizations to publish material to inform educators or the general public about important measurement issues. For example, in 1990 it partnered with the American Federation of Teachers (AFT) and the National Education Association (NEA) to produce the *Standards for Teacher Competence in the Educational Assessment of Students*. It has also been an active member on the Joint Committee on Testing Practices (JCTP) that produced the *ABCs of Testing,* which is a video and booklet designed to inform parents and other lay audiences about the use of tests in schools and about important characteristics of quality educational assessments. NCME also worked with JCTP to produce the *Code of Fair Testing Practices,* which describes the responsibilities test developers and test users have for ensuring fair and appropriate testing practices. NCME disseminates this document for free at its website. NCME also publishes a quarterly newsletter, which can also be downloaded for free from its website.

## Annual Conference

In addition to the aforementioned publications, the NCME's annual conference is another mechanism

with which it helps disseminate new findings and research on educational measurement. The annual conference, typically held in March or April, features three full days of paper sessions, symposia, invited speakers, and poster sessions in which psychometricians and other measurement practitioners can learn and dialog about new developments and issues in educational measurement and research. About 1,200 members attend the conference each year.

### Governance Structure

The governance structure of NCME consists of an Executive Committee (President, President-Elect, and Past President) and a six-member Board of Directors. The Board of Directors includes all elected positions, and including the President-Elect (also referred to as Vice President). In addition, NCME has 20 volunteer committees that are run by its members. Examples of these committees include the Outreach and Partnership Committee and the Diversity Issues and Testing Committee.

### Joining NCME

NCME is open to all professionals and membership includes subscriptions to the NCME Newsletter, JEM, and EM:IP. Graduate students can join for a reduced rate and can receive all three publications as part of their membership. All professionals interested in staying current with respect to new developments and research related to assessing students are encouraged to become members. To join the NCME visit the NCME website or write the NCME Central Office at 2810 Crossroads Drive, Suite 3800, Madison WI, 53718.

*Stephen G. Sireci*

***See also*** American Educational Research Association; American Statistical Association; "On the Theory of Scales of Measurement"

#### Further Readings

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

American Federation of Teachers, National Education Association, & National Council on Measurement in Education. (1990). Standards for teacher competence in the educational assessment of students. *Educational Measurement: Issues and Practice 9,* 30–34.

Brennan, R. L. (Ed). (2006). *Educational measurement* (4th ed.). Washington, DC: American Council on Education/Praeger.

Coffman, W. (1989). *Past presidents' committee: A look at the past, present, and future of NCME. 1. A look at the past.* (ERIC Document Reproduction Service No. ED308242).

Joint Committee on Testing Practices. (1988). *Code of fair testing practices in education.* Washington, DC: American Psychological Association.

Joint Committee on Testing Practices. (1993). *ABCs of testing.* Washington, DC: National Council on Measurement in Education.

Joint Committee on Testing Practices. (2004). *Code of fair testing practices in education.* Washington, DC: American Psychological Association.

Lindquist, E. F. (Ed.). (1951). *Educational measurement.* Washington, DC: American Council on Education.

Linn, R. L. (Ed.). (1989). *Educational measurement* (3rd ed.). Washington, DC: American Council on Education.

Millman, J. (1981). *Handbook of teacher evaluation.* Beverly Hills, CA: Sage.

Millman, J., & Darling-Hammond, L. (1990). *The new handbook of teacher evaluation: Assessing elementary and secondary school teachers.* Newbury Park, CA: Sage.

Thorndike, R. L. (Ed.). (1971). *Educational measurement* (2nd ed.). Washington, DC: American Council on Education.

#### Websites

National Council on Measurement in Education: http://www.ncme.org

## Natural Experiments

Natural experiments are designs that occur in nature and permit a test of an otherwise untestable hypothesis and thereby provide leverage to disentangle variables or processes that would otherwise be inherently confounded. Experiments in nature do not, by definition, have the sort of leverage that traditional experiments have because they were not manufactured to precise methodological detail;

they are fortuitous. They do, however, have distinct advantages over observational studies and might, in some circumstances, address questions that randomized controlled trials could not address. A key feature of natural experiments is that they offer insight into causal processes, which is one reason why they have an established role in developmental science.

Natural experiments represent an important research tool because of the methodological limits of naturalistic and experimental designs and the need to triangulate and confirm findings across multiple research designs. Notwithstanding their own set of practical limitations and threats to generalizability of the results, natural experiments have the potential to deconfound alternative models and accounts and thereby contribute significantly to developmental science and other areas of research. This entry discusses natural experiments in the context of other research designs and then illustrates how their use in developmental science has provided information about the relationship between early exposure to stress and children's development.

## The Scientific Context of Natural Experiments

The value of natural experiments is best appreciated when viewed in the context of other designs. A brief discussion of other designs is therefore illustrative. Observational or naturalistic studies—cross-sectional or longitudinal assessments in which individuals are observed and no experimental influence is brought to bear on them—generally cannot address causal claims. That is because a range of methodological threats, including selection biases and coincidental or spurious associations, undermine causal claims. So, for example, in the developmental and clinical psychology literature, there is considerable interest in understanding the impact of parental mental health—maternal depression is probably the most studied example—on children's physical and mental development. Dozens of studies have addressed this question using a variety of samples and measures. However, almost none of these studies—even large-scale cohort and population studies—are equipped to identify causal mechanisms for several reasons, including (a) genetic transmission is confounded with family processes and other psychosocial risks; (b) maternal depression is virtually always accompanied by other risks that are also reliably linked with children's maladjustment, such as poor parenting and marital conflict; and (c) mate selection for psychiatric disorders means that depressed mothers are more likely to have a partner with a mental illness, which confounds any specific "effect" that investigators might wish to attribute to maternal depression per se. Most of the major risk factors relevant to psychological well-being and public health co-occur; in general terms, risk exposures are not distributed randomly in the population. Indeed, one of the more useful lessons from developmental science has been to demonstrate the ways in which exposures to risk accrue in development.

One response to the problems in selection bias or confounded risk exposure is to address the problem analytically. That is, even if, for example, maternal depression is inherently linked with compromised parenting and family conflict, the "effect" of maternal depression might nevertheless be derived if the confounded variables (compromised parenting and family conflict) are statistically controlled for. There are some problems with that solution, however. If risk processes are confounded in nature, then statistical controlling for one or the other is not a satisfying solution; interpretations of the maternal depression "effect" will be possible but probably not (ecologically) valid. Sampling design strategies to obtain the same kind of leverage, such as sampling families with depressed mothers only if there is an absence of family conflict, will yield an unrepresentative sample of affected families with minimal generalizability. Case-control designs try to gain some leverage over cohort observational studies by tracking a group or groups of individuals, some of whom have a condition(s) of interest. Differences between groups are inferred to be attributable to the condition(s) of interest because the groups were matched on key factors. That is not always possible and the relevant factors to control for are not always known; as a result, between-group and even within-subject variation in these designs is subject to confounders.

A potential methodological solution is offered by experimental designs. So, for example, testing the maternal depression hypothesis referred to previously might be possible to the extent that some

affected mothers are randomly assigned to treatment for depression. That would offer greater purchase on the question of whether maternal depression per se was a causal contributor to children's adjustment difficulties. Interestingly, intervention studies have shown that there are a great many questions about causal processes that emerge even after a successful trial. For example, cognitive-behavioral treatment might successfully resolve maternal depression and, as a result, children of the treated mothers might show improved outcomes relative to children whose depressed mothers were not treated. It would not necessarily follow, however, that altering maternal depression was the causal mediator affecting child behavior. It might be that children's behavior improved because the no-longer-depressed mothers could engage as parents in a more effective manner and there was a decrease in inter-parental conflict, or any of several other secondary effects of the depression treatment. In other words, questions about causal mechanisms are not necessarily resolved fully by experimental designs.

Investigators in applied settings are also aware that some contexts are simply not amenable to randomized control. School-based interventions sometimes hit resistance to random assignment because principals, teachers, or parents object to the idea that some children needing intervention might not get the presumed better treatment. Court systems are often nonreceptive experimental proving grounds. That is, no matter how compelling data from a randomized control trial might be to address a particular question, there are circumstances in which a randomized control trial is extremely impractical or unethical.

Natural experiments are, therefore, particularly valuable where traditional nonexperimental designs might not be scientifically inadequate or where experimental designs may be practical or ethical. And, natural experiments are useful scientific tools even where other designs might be judged as capable of testing the hypothesis of interest. That is because of the need for findings to be confirmed not only by multiple studies but also by multiple designs. That is, natural experiments can provide a helpful additional scientific "check" on findings generated from naturalistic or experimental studies. There are many illustrations of the problems in relying on findings from a single design. Researchers are now accustomed to defining an effect or association as robust if it is replicated across samples and measures. Also, a finding should replicate across design. No single research sample and no single research design is satisfactory for testing causal hypotheses or inferring causal mechanisms.

Finally, identifying natural experiments can be an engaging and creative process, and studies based on natural experiments are far less expensive and arduous to investigate—they occur naturally—than those using conventional research designs; they can also be common. Thus, dramatic shifts in income might be exploited to investigate income dynamics and children's well-being; cohort changes in the rates of specific risks (e.g., divorce) might be used to examine psychosocial accounts for children's adjustment problems. Hypotheses about genetic and/or psychosocial risk exposure might be addressed using adoption and twin designs, and many studies exploit the arbitrariness of age cut-off for school to contrast exposure with maturation accounts of reading, language, and mathematic ability; the impact of compulsory schooling; and many other practical and conceptual questions.

Like all other forms of research design, natural experiments have their own special set of limitations. But, they can offer both novel and confirmatory findings. Examples of how natural experiments have informed the debate on early risk exposure and children's development are reviewed below.

## Natural Experiments to Examine the Long-Term Effects of Early Risk Exposure

Understanding the degree to which, and by what mechanisms, early exposure to stress has long-term effects is a primary question for developmental science with far-reaching clinical and policy applications. This area of inquiry has been extensively and experimentally studied in animal models. But animal studies are inadequate for deriving clinical and public health meaning; research in humans is essential. However, sound investigation to inform the debate in humans has been overshadowed by claims that might overplay the evidence, as in the case of extending animal findings to humans

willy-nilly. The situation is compounded by the general lack of relevant human studies that have leverage for deriving claims about early experience and exposure per se. That is, despite the hundreds of studies that assess children's exposure to early risk, almost none can differentiate the effects of early risk exposure from later risk exposure because the exposure to risk—maltreatment, poverty, and parental mental illness—is continuous rather than precisely timed or specific to the child's early life. Intervention studies have played an important role in this debate, and many studies now show long-term effects of early interventions. In contrast, because developmental timing was not separated from intervention intensity in most cases, these studies do not resolve issues about early experience as such. In other words, conventional research designs have not had much success in tackling major questions about early experience. It is not surprising, then, that natural experiments have played such a central role in this line of investigation.

Several different forms of natural experiments to study the effects of early exposure have been reported. One important line of inquiry is from the Dutch birth cohort exposed to prenatal famine during the Nazi blockade. Alan S. Brown and colleagues found that the rate of adult unipolar and bipolar depression requiring hospitalization was increased among those whose mothers experienced starvation during the second and third trimesters of pregnancy. The ability of the study to contrast rates of disorder among individuals whose mothers were and were not pregnant during the famine allowed unprecedented experimental "control" on timing of exposure. A second feature that makes the study a natural experiment is that it capitalized on a situation that is ethically unacceptable and so impossible to design on purpose.

Another line of study that has informed the early experience debate concerns individuals whose caregiving experience undergoes a radical change—far more radical than any traditional psychological intervention could create. Of course, radical changes in caregiving do not happen ordinarily. A notable exception is those children who are removed from abusive homes and placed into nonabusive or therapeutic settings (e.g., foster care). Studies of children in foster care are, therefore, significant because this is a population for whom the long-term outcomes are generally poor and because this is context in which natural experiments of altering care are conducted. Clearly, there are inherent complications, but the findings have provided some of the most interesting data in clinical and developmental psychology.

An even more extreme context involves children who experienced gross deprivation via institutional care and were then adopted into low-/normal-risk homes. There are many studies of this sort. One is the English and Romanian Adoptees (ERA) study, which is a long-term follow-up of children who were adopted into England after institutional rearing in Romanian; the study also includes an early adopted sample of children in England as a comparison group. A major feature of this particular natural experiment—and what makes it and similar studies of exinstitutionalized children noteworthy—is that there was a remarkable discontinuity in caregiving experience, from the most severe to a normal-risk setting. That feature offers unparallel leverage for testing the hypothesis that it is *early* caregiving risk that has persisting effects on long-term development. The success of the natural experiment design depends on many considerations, including the representativeness of the families who adopted from Romania to the general population of families, for example. A full account of this issue is not within the scope of this entry, but it is clear that the impact of findings from natural experiments needs to be judged in relation to the kinds of sampling and other methodological features.

The findings from studies of exinstitutionalized samples correspond across studies. So, for example, there is little doubt now from long-term follow-up assessments that early caregiving deprivation can have long-term impact on attachment and intellectual development, with a sizable minority of children showing persisting deficits many years after the removal from the institutional setting and despite many years in a resourceful, caring home environment. Findings also show that individual differences in response to early severe deprivation are substantial and just as continuous. Research into the effects of early experience has depended on these natural experiments because conventional research designs were either impractical or unethical.

*Thomas G. O'Connor*

*See also* Case-Only Design; Case Study; Narrative Research; Observational Research; Observations

**Further Readings**

Anderson, G. L., Limacher, M., Assaf, A. R., Bassford, T., Beresford, S. A., Black, H., et al. (2004). Effects of conjugated equine estrogen in postmenopausal women with hysterectomy: The women's health initiative randomized controlled trial. *Journal of the American Medical Association, 291,* 1701–1712.

Beckett, C., Maughan, B., Rutter, M., Castle, J., Colvert, E., Groothues, C., et al. (2006). Do the effects of early severe deprivation on cognition persist into early adolescence? Findings from the English and Romanian adoptees study. *Child Development,* 77, 696–711.

Campbell, F. A., Pungello, E. P., Johnson, S. M., Burchinal, M., & Ramey, C. T. (2001). The development of cognitive and academic abilities: Growth curves from an early childhood educational experiment. *Developmental Psychology, 37,* 231–242.

Collishaw, S., Goodman, R., Pickles, A., & Maughan, B. (2007). Modelling the contribution of changes in family life to time trends in adolescent conduct problems. *Social Science and Medicine, 65,* 2576–2587.

Grady, D., Rubin, S. M., Petitti, D. B., Fox, C. S., Black, D., Ettinger, B., et al. (1992). Hormone therapy to prevent disease and prolong life in postmenopausal women. *Annals of Internal Medicine,* 117, 1016–1037.

O'Connor, T. G., Caspi, A., DeFries, J. C., & Plomin, R. (2000). Are associations between parental divorce and children's adjustment genetically mediated? An adoption study. *Developmental Psychology, 36,* 429–437.

Tizard, B., & Rees, J. (1975). The effect of early institutional rearing on the behavioral problems and affectional relationships of four-year-old children. *Journal of Child Psychology and Psychiatry, 16,* 61–73.

# Naturalistic Inquiry

Naturalistic inquiry is an approach to understanding the social world in which the researcher observes, describes, and interprets the experiences and actions of specific people and groups in societal and cultural context. It is a research tradition that encompasses qualitative research methods originally developed in anthropology and sociology, including participant observation, direct observation, ethnographic methods, case studies, grounded theory, unobtrusive methods, and field research methods. Working in the places where people live and work, naturalistic researchers draw on observations, interviews, and other sources of descriptive data, as well as their own subjective experiences, to create rich, evocative descriptions and interpretations of social phenomena. Naturalistic inquiry designs are valuable for exploratory research, particularly when relevant theoretical frameworks are not available or when little is known about the people to be investigated. The characteristics, methods, indicators of quality, philosophical foundations, history, disadvantages, and advantages of naturalistic research designs are described below.

## Characteristics of Naturalistic Research

Naturalistic inquiry involves the study of a single case, usually a self-identified group or community. Self-identified group members are conscious of boundaries that set them apart from others. When qualitative (naturalistic) researchers select a case for study, they do so because it is of interest in its own right. The aim is not to find a representative case from which to generalize findings to other, similar individuals or groups. It is to develop interpretations and local theories that afford deep insights into the human experience.

Naturalistic inquiry is conducted in the field, within communities, homes, schools, churches, hospitals, public agencies, businesses, and other settings. Naturalistic researchers spend large amounts of time interacting directly with participants. The researcher is the research instrument, engaging in daily activities and conversations with group members to understand their experiences and points of view. Within this tradition, language is considered a key source of insight into socially constructed worlds. Researchers record participants' words and actions in detail with minimal interpretation. Although focused on words, narratives, and discourse, naturalistic researchers learn through all of their senses. They collect data at the following experiential levels: cognitive, social, affective, physical, and political/ideological. This strategy adds

depth and texture to the body of data qualitative researchers describe, analyze, and interpret.

Naturalistic researchers study research problems and questions that are initially stated broadly then gradually narrowed during the course of the study. In non-naturalistic, experimental research designs, terms are defined, research hypotheses stated, and procedures for data collection established in advance before the study begins. In contrast, qualitative research designs develop over time as researchers formulate new understandings and refine their research questions. Throughout the research process, naturalistic researchers modify their methodological strategies to obtain the kinds of data required to shed light on more focused or intriguing questions. One goal of naturalistic inquiry is to generate new questions that will lead to improved observations and interpretations, which will in turn foster the formulation of still better questions. The process is circular but ends when the researcher has created an account that seems to capture and make sense of all the data at hand.

## Naturalistic Research Methods

### *General Process*

When naturalistic researchers conduct field research, they typically go through the following common sequence of steps:

1. Gaining access to and entering the field site
2. Gathering data
3. Ensuring accuracy and trustworthiness (verifying and cross-checking findings)
4. Analyzing data (begins almost immediately and continues throughout the study)
5. Formulating interpretations (also an ongoing process)
6. Writing up findings
7. Member checking (sharing conclusions and conferring with participants)
8. Leaving the field site

### *Sampling*

Naturalistic researchers employ purposive rather than representative or random sampling methods. Participants are selected based on the purpose of the study and the questions under investigation, which are refined as the study proceeds. This strategy might increase the possibility that unusual cases will be identified and included in the study. Purposive sampling supports the development of theories grounded in empirical data tied to specific local settings.

### *Analyzing and Interpreting Data*

The first step in qualitative data analysis involves transforming experiences, conversations, and observations into text (data). When naturalistic researchers analyze data, they review field notes, interview transcripts, journals, summaries, and other documents looking for repeated patterns (words, phrases, actions, or events) that are salient by virtue of their frequency. In some instances, the researcher might use descriptive statistics to identify and represent these patterns.

Interpretation refers to making sense of what these patterns or themes might mean, developing explanations, and making connections between the data and relevant studies or theoretical frameworks. For example, reasoning by analogy, researchers might note parallels between athletic events and anthropological descriptions of ritual processes. Naturalistic researchers draw on their own understanding of social, psychological, and economic theory as they formulate accounts of their findings. They work inductively, from the ground up, and eventually develop location-specific theories or accounts based on analysis of primary data.

As a by-product of this process, new research questions emerge. Whereas traditional researchers establish hypotheses prior to the start of their studies, qualitative researchers formulate broad research questions or problem statements at the start, then reformulate or develop new questions as the study proceeds. The terms *grounded theory, inductive analysis,* and c*ontent analysis,* although not synonymous, refer to this process of making sense of and interpreting data.

## Evaluating Quality

The standards used to evaluate the adequacy of traditional, quantitative studies should not be used to assess naturalistic research projects. Quantitative

and qualitative researchers work within distinct traditions that rest on different philosophical assumptions, employ different methods, and produce different products. Qualitative researchers argue among themselves about how best to evaluate naturalistic inquiry projects, and there is little consensus on whether it is possible or appropriate to establish common standards by which such studies might be judged. However, many characteristics are widely considered to be indicators of merit in the design of naturalistic inquiry projects.

### *Immersion*

Good qualitative studies are time consuming. Researchers must become well acquainted with the field site and its inhabitants as well as the wider context within which the site is located. They also immerse themselves in the data analysis process, through which they read, review, and summarize their data.

### *Transparency and Rigor*

When writing up qualitative research projects, researchers must put themselves in the text, describing how the work was conducted, how they interacted with participants, how and why they decided to proceed as they did, and noting how participants might have been affected by these interactions. Whether the focus is on interview transcripts, visual materials, or field research notes, the analytical process requires meticulous attention to detail and an inductive, bottom-up process of reasoning that should be made clear to the reader.

### *Reflexivity*

Naturalistic inquirers do not seek to attain objectivity, but they must find ways to articulate and manage their subjective experiences. Evidence of one or more forms of reflexivity is expected in naturalistic inquiry projects. Positional reflexivity calls on researchers to attend to their personal experiences—past and present—and describe how their own personal characteristics (power, gender, ethnicity, and other intangibles) played a part in their interactions with and understandings of participants. Textual reflexivity involves skeptical, self-critical consideration of how authors (and the professional communities in which they work) employ language to construct their representations of the social world. A third form of reflexivity examines how participants and the researchers who study them create social order through practical, goal-oriented actions and discourse.

### *Comprehensiveness and Scope*

The cultural anthropologist Clifford Geertz used the term *thick description* to convey the level of rich detail typical of qualitative, ethnographic descriptions. When writing qualitative research reports, researchers place the study site and findings as a whole within societal and cultural contexts. Effective reports also incorporate multiple perspectives, including perspectives of participants from all walks of life (for example) within a single community or organization.

### *Accuracy*

Researchers are expected to describe the steps taken to verify findings and interpretations. Strategies for verification include triangulation (using and confirming congruence among multiple sources of information), member checking (negotiating conclusions with participants), and auditing (critical review of the research design, processes, and conclusions by an expert).

### *Claims and Warrants*

In well-designed studies, naturalistic researchers ensure that their conclusions are supported by empirical evidence. Furthermore, they recognize that their conclusions follow logically from the design of the study, including the review of pertinent literature, data collection, analysis, interpretation, and the researcher's inferential process.

### *Attention to Ethics*

Researchers should describe the steps taken to protect participants from harm and discuss any ethical issues that arose during the course of the study.

### *Fair Return*

Naturalistic inquiry projects are time consuming not only for researchers but also for participants,

who teach researchers about their ways of life and share their perspectives as interviewees. Researchers should describe what steps they took to compensate or provide fair return to participants for their help. Research leads to concrete benefits for researchers (degree completion or career advancement). Researchers must examine what benefits participants will gain as a result of the work and design their studies to ensure reciprocity (balanced rewards).

### *Coherence*

Good studies call for well-written and compelling research reports. Standards for writing are genre specific. Postmodern authors defy tradition through experimentation and deliberate violations of writing conventions. For example, some authors avoid writing in clear, straightforward prose to express more accurately the complexities inherent in the social world and within the representational process.

### *Veracity*

A good qualitative report brings the setting and its residents to life. Readers who have worked or lived in similar settings find the report credible because it reflects aspects of their own experiences.

### *Illumination*

Good naturalistic studies go beyond mere description to offer new insights into social and psychological phenomena. Readers should learn something new and important about the social world and the people studied, and they might also gain a deeper understanding of their own ways of life.

## Philosophical Foundations

Traditional scientific methods rest on philosophical assumptions associated with logical positivism. When working within this framework, researchers formulate hypotheses that are drawn from established theoretical frameworks, define variables by stipulating the processes used to measure them, collect data to test their hypotheses, and report their findings objectively. Objectivity is attained through separation of the researcher from participants and by dispassionate analysis and interpretation of results. In contrast, naturalistic researchers tap into their own subjective experiences as a source of data, seeking experiences that will afford them an intuitive understanding of social phenomena through empathy and subjectivity. Qualitative researchers use their subjective experiences as a source of data to be carefully described, analyzed, and shared with those who read their research reports.

For the naturalistic inquirer, objectivity and detachment are neither possible nor desirable. Human experiences are invariably influenced by the methods used to study them. The process of being studied affects all humans who become subjects of scientific attention. The presence of an observer affects those observed. Furthermore, the observer is changed through engaging with and observing the other. Objectivity is always a matter of degree.

Qualitative researchers are typically far less concerned about objectivity as this term is understood within traditional research approaches than with intersubjectivity. Intersubjectivity is the process by which humans share common experiences and subscribe to shared understandings of reality. Naturalistic researchers seek involvement and engagement rather than detachment and distance. They believe that humans are not rational beings and cannot be understood adequately through objective, disembodied analysis. Authors critically examine how their theoretical assumptions, personal histories, and methodological decisions might have influenced findings and interpretations (positional reflexivity). In a related vein, naturalistic researchers do not believe that political neutrality is possible or helpful. Within some qualitative research traditions, researchers collaborate with participants to bring about community-based political and economic change (social justice).

Qualitative researchers reject determinism, the idea that human behaviors are lawful and can be predicted. Traditional scientists try to discover relationships among variables that remain consistent across individuals beyond the experimental setting. Naturalistic inquiry rests on the belief that studying humans requires different methods than those used to study the material world. Advocates emphasize that no shared, universal reality remains

constant over time and across cultural groups. The phenomena of most interest to naturalistic researchers are socially constructed, constantly changing, and multiple. Naturalistic researchers hold that all human phenomena occur within particular contexts and cannot be interpreted or understood apart from these contexts.

## History

The principles that guide naturalistic research methods were developed in biology, anthropology, and sociology. Biologist Charles Darwin developed the natural history method, which employs detailed observation of the natural world directed by specific research questions and theory building based on analysis of patterns in the data, followed by confirmation (testing) with additional observations in the field. Qualitative researchers use similar strategies, which transform experiential, qualitative information gathered in the field into data amenable to systematic investigation, analysis, and theory development.

Ancient adventurers, writers, and missionaries wrote the first naturalistic accounts, describing the exotic people they encountered on their travels. During the early decades of the 20th century, cultural anthropologists and sociologists pioneered the use of ethnographic research methods for the scientific study of social phenomena. Ethnography is both a naturalistic research methodology and a written report that describes field study findings. Although there are many different ethnographic genres, all of them employ direct observation of naturally occurring events in the field. Early in the 20th century, University of Chicago sociologists used ethnographic methods to study urban life, producing pioneering studies of immigrants, crime, work, youth, and group relations. Sociologist Herbert Blumer, drawing on George Herbert Mead, William I. Thomas, and John Dewey, developed a rationale for the naturalistic study of the social world. In the 1970s, social scientists articulated ideas and theoretical issues pertinent to naturalistic inquiry. Interest in qualitative research methods grew. In the mid-1980s, Yvonne Lincoln and Egon Guba published *Naturalistic Inquiry,* which provided a detailed critique of positivism and examined implications for social research. Highlighting the features that set qualitative research apart from other methods, these authors also translated key concepts across what they thought were profoundly different paradigms (disciplinary worldviews). In recent years, qualitative researchers considered the implications of critical, feminist, postmodern, and poststructural theories for their enterprise. The recognition or rediscovery that researchers create the phenomena they study and that language plays an important part in this process has inspired methodological innovations and lively discussions. The discourse on naturalistic inquiry remains complex and ever changing. New issues and controversies emerge every year, reflecting philosophical debates within and across many academic fields.

## Methodological Disadvantages and Advantages

### *Disadvantages*

Many areas are not suited to naturalistic investigation. Naturalistic research designs cannot uncover cause and effect relationships and they cannot help researchers evaluate the effectiveness of specific medical treatments, school curricula, or parenting styles. They do not allow researchers to measure particular attributes (motivation, reading ability, or test anxiety) or to predict the outcomes of interventions with any degree of precision. Qualitative research permits only claims about the specific case under study. Generalizations beyond the research site are not appropriate. Furthermore, naturalistic researchers cannot set up logical conditions whereby they can demonstrate their own assumptions to be false.

Naturalistic inquiry is time consuming and difficult. Qualitative methods might seem to be easier to use than traditional experimental and survey methods because they do not require mastery of technical statistical and analytical methods. However, naturalistic inquiry is one of the most challenging research approaches to learn and employ. Qualitative researchers tailor methods to suit each project, revising data-collection strategies as questions and research foci emerge. Naturalistic researchers must have a high tolerance for uncertainty and the ability to work independently for extended periods of time, and these researchers must also be able to think creatively under pressure.

### *Advantages*

Once controversial, naturalistic research methods are now used in social psychology, developmental psychology, qualitative sociology, and anthropology. Researchers in professional schools (education, nursing, health sciences, law, social work, and counseling) and applied fields (regional planning, library science, program evaluation, information science, and sports administration) employ naturalistic strategies to investigate social phenomena. Naturalistic approaches are well suited to the study of groups about which little is known. They are also holistic and comprehensive. Qualitative researchers try to tell the whole story, in context. A well-written report has some of the same characteristics as a good novel, bringing the lives of participants to life. Naturalistic methods help researchers understand how people view the world, what they value, and how these values and cognitive schemas are reflected in practices and social structures. Through the study of groups unlike their own, researchers learn that many different ways are available to raise children, teach, heal, maintain social order, and initiate change. Readers learn about the extraordinary variety of designs for living and adaptive strategies humans have created, thus broadening awareness of possibilities beyond conventional ways of life. Thus, naturalistic inquiry can provide insights that deepen our understanding of the human experience and generate new theoretical insights. For researchers, the process of performing qualitative research extends and intensifies the senses and provides interesting and gratifying experiences as relationships are formed with the participants from whom and with whom one learns.

*Jan Armstrong*

***See also*** Ethnography; Grounded Theory; Interviewing; Naturalistic Observation; Qualitative Research

#### Further Readings

Blumer, H. (1969). *Symbolic interactionism: Perspective and method.* Englewood Cliffs, NJ: Prentice Hall.

Bogdan, R., & Bicklen, S. K. (2006). *Qualitative research for education: An introduction to theories and methods* (5th ed.). Boston: Allyn & Bacon.

Creswell, J. W. (2006). *Qualitative inquiry and research design: Choosing among five approaches* (2nd ed.). Thousand Oaks, CA: Sage.

Denzin, N. K., & Lincoln, Y. S. (2005). *Sage handbook of qualitative research* (3rd ed.). Thousand Oaks, CA: Sage.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry.* Beverly Hills, CA: Sage.

Macbeth, D. (2001). On "reflexivity" in qualitative research: Two readings, and a third. *Qualitative Inquiry, 7,* 35–68.

Marshall, C., & Rossman, G. B. (2006). *Designing qualitative research* (4th ed.). Thousand Oaks, CA: Sage.

Seale, C. (1999). Quality in qualitative research. *Qualitative Inquiry, 5,* 465–478.

## Naturalistic Observation

Naturalistic observation is a nonexperimental, primarily qualitative research method in which organisms are studied in their natural settings. Behaviors or other phenomena of interest are observed and recorded by the researcher, whose presence might be either known or unknown to the subjects. This approach falls within the broader category of field study, or research conducted outside the laboratory or institution of learning. No manipulation of the environment is involved in naturalistic observation, as the activities of interest are those manifested in everyday situations. This method is frequently employed during the initial stage of a research project, both for its wealth of descriptive value and as a foundation for hypotheses that might later be tested experimentally.

Zoologists, naturalists, and ethologists have long relied on naturalistic observation for a comprehensive picture of the variables coexisting with specific animal behaviors. Charles Darwin's 5-year voyage aboard the H.M.S. *Beagle,* which is an expedition that culminated in his theory of evolution and the publication of his book *On the Origin of Species* in 1859, is a paradigm of research based on this method. Studies of interactions within the social structures of primates by both Dian Fossey and Jane Goodall relied on observation in the subjects' native habitats. Konrad Lorenz, Niko Tinbergen, and Karl von Frisch advanced the

understanding of communication among animal species through naturalistic observation, introducing such terminology as imprinting, fixed action pattern, sign stimulus, and releaser to the scientific lexicon. All the investigations mentioned here were notable for their strong ecological validity, as they were conducted within a context reflective of the normal life experiences of the subjects. It is highly doubtful that the same richness of content could have been obtained in an artificial environment devoid of concurrent factors that would have normally accompanied the observed behaviors.

The instances in which naturalistic observation also yields valuable insight to psychologists, social scientists, anthropologists, ethnographers, and behavioral scientists in the study of human behavior are many. For example, social deficits symptomatic of certain psychological or developmental disorders (such as autism, childhood aggression, or anxiety) might be evidenced more clearly in a typical context than under simulated conditions. The dynamics within a marital or family relationship likewise tend to be most perceptible when the participants interact as they would under everyday circumstances. In the study of broader cultural phenomena, a researcher might collect data by living among the population of interest and witness-ing activities that could only be observed in a real-life situation after earning their trust and their acceptance as an "insider."

This entry begins with the historic origins of naturalistic observation. Next, the four types of naturalistic observation are described and naturalistic observation and experimental methods are compared. Last, this entry briefly discusses the future direction of naturalistic observation.

## Historic Origins

The field of qualitative research gained prominence in the United States during the early 20th century. Its emergence as a recognized method of scientific investigation was taking place simultaneously in Europe, although the literature generated by many of these proponents was not available in the Western Hemisphere until after World War II. At the University of Chicago, such eminent researchers as Robert Park, John Dewey, Margaret Mead, and Charles Cooley contributed greatly to the development of participant observation methodology in the 1920s and 1930s. The approach became widely adopted among anthropologists during these same two decades. In Mead's 1928 study "Coming of Age in Samoa," data were collected while she resided among the inhabitants of a small Samoan village, making possible her groundbreaking revelations on the lives of girls and women in this island society.

Over the years, naturalistic observation became a widely used technique throughout the many scientific disciplines concerned with human behavior. Among its best-known practitioners was Jean Piaget, who based his theory of cognitive development on observations of his own children throughout the various stages of their maturation; in addition, he would watch other children at play, listening to and recording their interactions. Jeremy Tunstall conducted a study of fishermen in the English seaport of Hull by living among them and working beside them, a sojourn that led to the publication of his book *Fishermen: The Sociology of an Extreme Occupation* in 1962. Stanley Milgram employed naturalistic observation in an investigation on the phenomenon of "familiar strangers" (people who encountered but never spoke to one another) among city dwellers, watching railway commuters day after day as they waited to board the train to their workplaces in New York City. At his Family Research Laboratory in Seattle, Washington, John Gottman has used audiovisual monitoring as a component of his marriage counseling program since 1986. Couples stay overnight in a fabricated apartment at the laboratory, and both qualitative data (such as verbal interactions, proxemics, and kinesics) and quantitative data (such as heart rate, pulse amplitude, and skin conductivity) are collected. In his book *The Seven Principles for Making Marriage Work,* Gottman reported that the evidence gathered during this phase of therapy enabled him to predict whether a marriage would fail or succeed with 91% accuracy.

## Types of Naturalistic Observation

Naturalistic observation might be divided into four distinct categories. Each differs from the others in terms of basic definitions, distinguishing features, strengths and limitations, and appropriateness for specific research designs.

### *Overt Participant Observation*

In this study design, subjects are aware that they are being observed and are apprised of the research purpose prior to data collection. The investigator participates in the activities of the subjects being studied and might do this by frequenting certain social venues, taking part in the affairs of an organization, or living as a member of a community. For example, a researcher might travel with a tour group to observe how people cope with inability to communicate in a country where an unfamiliar language is spoken.

The nature of this study design obviates several ethical concerns, as no deception is involved. However, reactivity to the presence of the observer might compromise internal validity. The Hawthorne effect, in which behavioral and performance-related changes (usually positive) occur as a result of the experimenter's attention, is one form of reactivity. Other problems might ensue if subjects change their behaviors after learning the purpose of the experiment. These artifacts might include social desirability bias, attempts to confirm any hypotheses stated or suggested by the investigator, or noncooperation with the aim of disconfirming the investigator's hypotheses. Reluctant subjects might also anticipate the presence of the investigator and might take measures to avoid being observed.

### *Covert Participant Observation*

In many circumstances, disclosure of the investigator's purpose would jeopardize the successful gathering of data. In such cases, covert participant observation allows direct involvement in the subjects' activities without revealing that a study is being conducted. The observer might join or pretend to join an organization; assume the role of a student, instructor, or supervisor; or otherwise mingle unobtrusively with subjects to gain access to relevant information. Observation via the internet might involve signing up for a website membership using a fictitious identity and rationale for interest in participation. A study in which investigators pretend to be cult members to gather information on the dynamics of indoctrination is one instance of covert participant observation.

Certain distinct benefits are associated with this approach. As subjects remain oblivious to the presence of the researcher, reactivity to the test situation is eliminated. Frequently, it is easier to win the confidence of subjects if they believe the researcher to be a peer. However, measures taken to maintain secrecy might also restrict the range of observation. Of even greater concern is the potential breach of ethics represented by involving subjects in a study without their informed consent. In all such cases, it is important to consider whether the deception involved is justified by the potential benefits to be reaped. The ethical principles of such professional organizations as the American Psychological Association, American Medical Association, and American Counseling Association stress the overarching goal of promoting the welfare and respecting the dignity of the client or patient. This can be summarized as an aspirational guideline to do no harm. Apart from avoiding physical injury, it is necessary to consider and obviate any aspects of the study design that could do psychological or emotional damage. Even when potentially detrimental effects fail to materialize, the practitioner is responsible for upholding the principle of integrity in his or her professional conduct and for performing research honestly and without misrepresentation. Finally, subjects might eventually learn of their involuntary inclusion in research, and the consequent potential for litigation cannot be taken lightly. The issue of whether to extenuate deception to gain information that might not otherwise be procured merits serious deliberation, and is not treated casually in research design.

### *Overt Nonparticipant Observation*

The primary distinction between this method and overt participant observation is the role of the observer, who remains separate from the subjects being studied. Of the two procedures, nonparticipant observation is implemented in research more frequently. Subjects acknowledge the study being conducted and the presence of the investigator, who observes and records data but does not mingle with the subjects. For instance, a researcher might stand at a subway station, watching commuters and surreptitiously noting the frequency of discourteous behaviors during rush hour as opposed to off-peak travel times.

A major advantage of overt nonparticipant observation is the investigator's freedom to use

various tools and instruments openly, thus enabling easier and more complete recording of observations. (In contrast, a covert observer might be forced to write hasty notes on pieces of paper to avoid suspicion and to attempt reconstruction of fine details from memory later on.) Still, artifacts associated with awareness of the investigator's presence might persist, even though observation from a distance might tend to exert less influence on subjects' behavior. In addition, there is virtually no opportunity to question subjects should the researcher wish to obtain subsequent clarification of the meaning attached to an event. The observer might, thus, commit the error of making subjective interpretations based on inconclusive evidence.

### *Covert Nonparticipant Observation*

This procedure involves observation conducted apart from the subjects being studied. As in covert participant observation, the identity of the investigator is not revealed. Data are often secretly recorded and hidden; alternatively, observations might be documented at a later time when the investigator is away from the subjects. Witnessing events by means of electronic devices is also a form of covert nonparticipant observation. For example, the researcher might watch a videotape of children at recess to observe peer aggression.

The covert nonparticipant observer enjoys the advantages of candid subject behavior as well as the availability of apparatus with which to record data immediately. However, as in covert participant observation, measures taken to preserve anonymity might also curtail access to the full range of observations. Remote surveillance might similarly offer only a limited glimpse of the sphere of contextual factors, thereby diminishing the usefulness of the data. Finally, the previously discussed ethical infractions associated with any form of covert observation, as well as the potential legal repercussions, make using this method highly controversial.

## Comparing Naturalistic Observation With Experimental Methods

The advantages offered by naturalistic observation are many, whether in conjunction with experimental research or as the primary constituent of a study. First, there is less formal planning than in the experimental method, and also more flexibility is involved in accommodating change throughout the research process. These attributes make for an ideal preliminary procedure, one that might serve to lay the groundwork for a more focused investigation. As mentioned earlier, unexpected observations might generate new hypotheses, thereby contributing to the comprehensiveness of any research based thereon.

By remaining unobtrusive, the observer has access to behaviors that are more characteristic, more spontaneous, and more diverse that those one might witness in a laboratory setting. In many instances, such events simply cannot be examined in a laboratory setting. To learn about the natural behavior of a wild animal species, the workplace dynamics of a corporate entity, or the culturally prescribed roles within an isolated society, the investigator must conduct observations in the subjects' day-to-day environment. This requirement ensures a greater degree of ecological validity than one could expect to achieve in a simulated environment. However, there are no implications for increased external validity. As subjects are observed by happenstance, not selected according to a sampling procedure, representativeness cannot be guaranteed. Any conclusions drawn must necessarily be limited to the sample studied and cannot generalize to the population.

There are other drawbacks to naturalistic observation vis-à-vis experimental methods. One of these is the inability to control the environment in which subjects are being observed. Consequently, the experimenter can derive descriptive data from observation but cannot establish cause-and-effect relationships. Not only does this preclude explanation of why behaviors occur, but also it limits the prediction of behaviors. Additionally, the natural conditions observed are unique in all instances, thus rendering replication unfeasible.

The potential for experimenter bias is also significant. Whereas the number of times a behavior is recorded and the duration of the episode are both unambiguous measures, the naturalistic observer lacks a clear-cut system for measuring the extent or magnitude of a behavior. Perception of events might thus be influenced by any number of factors, including personal worldview. An especially problematic situation might arise when the observer is informed of the hypothesis and of the conditions under

investigation, as this might lead to seeking confirmatory evidence. Another possible error is that of the observer recording data in an interpretative rather than a descriptive manner, which can result in an ex post facto conclusion of causality. The researcher's involvement with the group in participant observation might constitute an additional source of bias. Objectivity can suffer because of group influence, and data might also be colored by a strong positive or negative impression of the subjects.

Experimental approaches to research differ from naturalistic observation on a number of salient points. One primary advantage of the true experiment is that a hypothesis can be tested and a cause-and-effect relationship can be demonstrated. The independent variable of interest is systematically manipulated, and the effects of this manipulation on the dependent variable are observed. Because the researcher controls the environment in which the study is conducted, it is thus possible to eliminate confounding variables. Besides enabling attribution of causality, this design also provides evidence of why a behavior occurs and allows prediction of when and under what conditions the behavior is likely to occur again. Unlike naturalistic observation, an experimental study can possess internal and external validity, although the controls inherent in this approach can diminish ecological validity, as it might be difficult to eliminate extraneous variables while maintaining some semblance of a real-world setting.

An additional benefit of experimental research is the relative stability of the environment in which the researcher conducts the study. In contrast, participant observation might entail a high degree of stress and personal risk when working with certain groups (such as gang members or prison inmates). This method also demands investment of considerable time and expense, and the setting might not be conducive to management of other responsibilities.

Although experimental design is regarded as a more conclusive method than naturalistic observation and is more widely used in science, it is not suitable for all research. Ethical and legal guidelines might forbid an experimental treatment if it is judged capable of harming subjects. For example, in studying the progression of a viral infection such as HIV, the investigator is prohibited from causing subjects to contract the illness and must instead recruit those who have already tested positive for the virus. Similarly, only preexisting psychiatric conditions (such as posttraumatic stress disorder) are studied, as subjects cannot be exposed to manipulations that could cause psychological or emotional trauma. Certain factors might be difficult or impossible to measure, as is the case with various cognitive processes.

The researcher's choice of method can either contribute to or, conversely, erode scientific rigor. If a convenience sample is used, if there are too few subjects in the sample, if randomization is flawed, or if the sample is otherwise not representative of the population from which it is selected, then the study will not yield generalizable results. The use of an instrument with insufficient reliability and validity might similarly undermine the experimental design. Nonetheless, bias and human error are universal in all areas of research. Self-awareness, critical thinking, and meticulous research methods can do much to minimize their ill effects.

## Future Directions

Robert Elliott and his colleagues proposed new guidelines for the publication of qualitative research studies in 1999, with the goal of encouraging legitimization, quality control, and subsequent development of this approach. Their expoition of both the traditional value and the current evolution of qualitative research was a compelling argument in support of its function not only as a precursor to experimental investigations but also as a method that addressed a different category of questions and therefore merited recognition in its own right. Given the ongoing presence of nonexperimental approaches in college and university curricula and in the current literature, it is likely that naturalistic observation will continue to play a vital role in scientific research.

*Barbara M. Wells*

***See also*** Descriptive Statistics; Ecological Validity; Naturalistic Inquiry; Observational Research; Observations; Qualitative Research

### Further Readings

Davidson, B., Worrall, L., & Hickson, L. (2003). Identifying the communication activities of older

people with aphasia: Evidence from naturalistic observation. *Aphasiology,* 17, 243–264.
Education Forum. *Primary research methods.* Retrieved August 24, 2008, from http://www.educationforum.co.uk/Health/primarymethods.ppt
Elliott, R., Fisher, C. T., & Rennie, D. L. (1999). Evolving guidelines for publication of qualitative research studies in psychology and related fields. *British Journal of Clinical Psychology,* 38, 215–229.
Fernald, D. (1999). Research methods. In *Psychology.* Upper Saddle River, NJ: Prentice Hall.
Mehl, R. (2007). Eavesdropping on health: A naturalistic observation approach for social health research. *Social and Personality Psychology Compass,* 1, 359–380.
Messer, S. C., & Gross, A. M. (1995). Childhood depression and family interaction: A naturalistic observation study. *Journal of Clinical Child Psychology,* 24, 77–88.
Pepler, D. J., & Craig, W. M. (1995). A peek behind the fence: Naturalistic observations of aggressive children with remote audiovisual recording. *Developmental Psychology,* 31, 548–553.
Spata, A. V. (2003). *Research methods: Science and diversity.* New York: Wiley.
Weinrott, M. R., & Jones, R. R. (1984). Overt versus covert assessment of observer reliability. *Child Development,* 55, 1125–1137.

# Nested Factor Design

In nested factor design, two or more factors are not completely crossed; that is, the design does not include each possible combination of the levels of the factors. Rather, one or more factors are nested within the levels of another factor. For example, in a design in which a factor (factor B) has four levels and is nested within the two levels of a second factor (factor A), levels 1 and 2 of factor B would only occur in combination with level 1 of factor A and levels 3 and 4 of factor B would only be combined with level 2 of factor A. In other words, in a nested factor design, there are cells that are empty. In the described design, for example, no observations are made for the combination of level 1 of factor A and level 3 of factor B. When a factor B is nested under a factor A, this is denoted as B(A). In more complex designs, a factor can also be nested under combinations of other factors. A common example in which factors are nested is within treatments, for example, the evaluation of psychological treatments when therapists or treatment centers provide one treatment to more than one participant. Because each therapist or treatment center provides only one treatment, the provider or treatment center factor is nested under only one level of the treatment factor. Nested factor designs are also common in educational research in which classrooms of students are nested within classroom interventions. For example, researchers commonly assign whole classrooms to different levels of a classroom-intervention factor. Thus, each classroom, or cluster, is assigned to only one level of the intervention factor and is said to be nested under this factor. Ignoring a nested factor in the evaluation of a design can lead to consequences that are detrimental to the validity of statistical decisions. The main reason for this is that the observations within the levels of a nested factor are likely to not be independent of each other but related. The magnitude of this relationship can be expressed by a so-called intraclass correlation coefficient $\rho_I$.

The focus of this entry is on the most common nested design: the two-level nested design. This entry discusses whether nested factors are random or fixed effects and the implications of nested designs on statistical power. In addition, the criteria to determine which model to use and the consequences of ignoring nested factors are also examined.

## Two-Level Nested Factor Design

The most common nested design involves two factors with a factor B nested within the levels of a second factor A. The linear structural model for this design can be given as follows:

$$Y_{ijk} = \mu + \alpha_j + \beta_{k(j)} + \varepsilon_{ijk}, \qquad (1)$$

where $Y_{ijk}$ is the observation for the $i$th subject ($i = 1, 2, \ldots, n$) in the $j$th level of factor A ($j = 1, 2, \ldots, p$) and the $k$th level of the nested factor B ($k = 1, 2, \ldots, q$), $\mu$ is the grand mean, $\alpha_j$ is the effect for the $j$th treatment, $\beta_{k(j)}$ is the effect of the $k$th provider nested under the $j$th treatment, and $\varepsilon_{ijk}$ is the error of the observation (within cell variance). Note that because factors A and B are not completely crossed, the model does not include an interaction term because it cannot be estimated separately from the error term. More generally speaking, because nested factor designs have not as many cells as

completely crossed designs, one cannot perform all tests for main effects and interactions.

The assumptions of the nested model are that the effects of the fixed factor A sum up to zero

$$\sum_j \alpha_j = 0, \tag{2}$$

that errors are normally distributed and have an expected value of zero

$$\varepsilon_{ijk} \sim N\left(0, \sigma^2_{\varepsilon_{i(jk)}}\right), \text{ and that} \tag{3}$$

$$\varepsilon_{i(jk)}, \alpha_j, \text{ and } \beta_{k(j)} \text{ are pairwise independent.} \tag{4}$$

In the next section, the focus is on nested factor designs with two factors with one of the factors nested under the other. More complex models can be built analogously. For example, the model equation for a design with two crossed factors, A and B, and a third factor nested within factor C is described by the following structural model:

$$Y_{ijkm} = \mu + \alpha_i + \beta_j + \gamma_{k(i)} + (\alpha\beta)_{jk(i)} + \varepsilon_{ijkm}. \tag{5}$$

## Nested Factors as Random and Fixed Effects

In experimental and quasi-experimental designs, the factor under which the second factor is nested is almost always conceptualized as a fixed factor. That is, a researcher seeks to make inferences about the specific levels of the factor included in the study and does not consider them random samples from a population of levels. For example, if a researcher compares different treatment groups or sets of experimental stimuli with different characteristics, the researcher's goal is to make inferences about the specific treatments implemented in the study or the stimulus properties that are manipulated.

The assumptions concerning the nested factor depend on whether it is considered to be a random or a fixed effect. Although in nested designs the nested factor is often considered to be random, there is no necessity to do this. The question of whether a factor should be treated as random or fixed is generally answered as follows: If the levels of a nested factor that are included in a study are considered to be a random sample of an underlying population of levels and if it is intended to generalize to the levels of the factor that are not included in the study (the universe of levels), then a nested factor should be treated as a *random factor*. The assumption that the levels included in the study are representative of an underlying population requires that the levels are drawn at random from the universe of levels. Thus, nested factor levels are treated like subjects who are also considered to be random samples from the population of all subjects. The resulting model is called a *mixed model* and assumes that the effects of factor B are normally distributed with a mean of zero, specifically:

$$\beta_{k(j)} \sim N\left(0, \sigma^2_{\beta_{k(j)}}\right). \tag{6}$$

A nested factor is correctly conceptualized as a *fixed factor* if a researcher only seeks to make an inference about the specific levels of the nested factor included in his or her study and if the levels included in the study are not drawn at random from a universe of levels. For example, if a researcher wants to make an inference about the specific treatment centers (nested within different treatments) that are included in her study, the nested factor is correctly modeled as a fixed effect. The corresponding assumption of the *fixed model* is

$$\sum_{k(j)} \beta_{k(j)} = 0, \tag{7}$$

that is, the effects of the nested factor add up to zero within each level of factor A. The statistical conclusion is then conditional on the factor levels included in the study.

In both the mixed model and the fixed model, the variance for the total model is given by

$$\sigma^2_{\text{total}} = \sigma^2_{\text{A}} + \sigma^2_{\text{B}} + \sigma^2_{\text{within}}. \tag{8}$$

In the population model, the proportion of variance accounted for by factor A can be expressed as

$$\omega_2 = \frac{\sigma^2_{\text{A}}}{\sigma^2_{\text{total}}} = \frac{\sigma^2_{\text{A}}}{\sigma^2_{\text{A}} + \sigma^2_{\text{B}} + \sigma^2_{\text{within}}}, \tag{9}$$

that is, the variance caused by factor A divided by the total variance (i.e., the sum of the variance caused by the treatments, variance caused by providers, and within-cell variance). Alternatively, the

treatment effect can be expressed as the partial effect size

$$\omega^2_{\text{part}} = \frac{\sigma^2_{\text{A}}}{\sigma^2_{\text{A}} + \sigma^2_{\text{within}}}, \tag{10}$$

that is, the variance caused by factor A divided by the sum of the variance caused by the treatments and within-cell variance. The partial effect size reflects the effectiveness of the factor A independent of additional effects of the nested factor B.

The effect resulting from the nested factor can be analogously defined as partial effect size, as follows:

$$\omega^2_{\text{B}} = \frac{\sigma^2_{\text{B}}}{\sigma^2_{\text{B}} + \sigma^2_{\text{within}}} = \rho_I. \tag{11}$$

If the nested factor is modeled as random, this effect is equal to the intraclass correlation coefficient $\rho_I$. This means that intraclass correlations $\rho_I$ are partial effect sizes of the nested factor B (i.e., independent of the effects of factor A). The intraclass correlation coefficient, which represents the relative amount of variation attributable to the nested factor, is also a measure of the similarity of the observations within the levels of the nested factors. It is, therefore, a measure of the degree to which the assumption of independence—required if the nested factor is ignored in the analysis and individual observations are the unit of analysis—is violated. Ignoring the nested factor if the intraclass correlation is not zero can lead to serious problems, especially to alpha inflation.

### *Sample Statistics*

The source tables for the mixed model (factor A fixed and factor B random) and the fixed model (both factors fixed) are presented in Table 1.

The main difference between the mixed and the fixed model is that in the mixed model, the expected mean square for factor A contains a term that includes the variance caused by the nested factor B (viz., $n\sigma^2_{\text{B}}$), whereas in the fixed model, the expected mean square for treatment effects contains no such term. Consequently, in the mixed-model case, the correct denominator to calculate the test statistic for factor A is the mean square for the nested factor, namely

$$F_{\text{mixed}} = \frac{MS_{\text{A}}}{MS_{\text{B}}}. \tag{12}$$

Note that the degrees of freedom of the denominator are exclusively a function of the number of levels of the factors A and B and are not influenced by the number of subjects within each cell of the design.

In the fixed-model case, the correct denominator is the mean square for within cell variation, namely

$$F_{\text{fixed}} = \frac{MS_{\text{A}}}{MS_{\text{within}}}. \tag{13}$$

Note that the within-cell variation does not include the variation resulting from the nested factor and that the degrees of freedom of the denominator are largely determined by the number of subjects.

The different ways the tests statistics for the non-nested factor A are calculated reflect the different underlying model assumptions. In the mixed model, levels of the nested factor are treated as a random sample from an underlying universe of levels. Because variation caused by the levels of the nested factor sampled in a particular study will randomly vary across repetitions of a study, this variation is considered to be error. In the fixed model, it is assumed that the levels of the nested

**Table 1** Sources of Variance and Expected Mean Squares for Nested Design: Factor A Fixed and Nested Factor B Random (Mixed Model) Versus Factor A Fixed and Nested Factor B Fixed (Fixed Model)

| *Source* | *SS* | *df* | *MS* | *E(MS) Mixed Model* | *E(MS) Fixed Model* |
|---|---|---|---|---|---|
| A | $SS_{\text{A}}$ | $p-1$ | $\frac{SS_{\text{A}}}{p-1}$ | $\sigma^2_w + n\sigma^2_{\text{B}} + [p/(p-1)]nq\sigma^2_{\text{A}}$ | $\sigma^2_w + [p/(p-1)]nq\sigma^2_{\text{A}}$ |
| B(A) | $SS_{\text{B}}$ | $p(q-1)$ | $\frac{SS_{\text{B}}}{p(q-1)}$ | $\sigma^2_w + n\sigma^2_{\text{B}}$ | $\sigma^2_w + [q/(q-1)]n\sigma^2_{\text{B}}$ |
| Within cell ($w$) | $SS_w$ | $pq(n-1)$ | $\frac{SS_{\text{within}}}{pq(n-1)}$ | $\sigma^2_w$ | $\sigma^2_w$ |

*Notes:* The number of levels of factor A is represented by $p$, the number of levels of the nested factor B within each level of factor A is represented by $q$, and the number of subjects within each level of factor B is represented by $n$. $df$ = degrees of freedom; $SS$ = sum of squares; $MS$ = mean square; E($MS$) = expected mean square; A = factor A; B = nested factor B; w = within cell.

factor included in a particular study will not vary across replications of a study, and variation from the nested factor is removed from the estimated error.

### *Effect Size Estimates in Nested Factor Designs*

The mixed and the fixed models vary with respect to how population effect sizes are estimated. First, population effects are typically not estimated by the sample effect size, namely

$$\eta^2_{\mathrm{A}} = \frac{SS_{\mathrm{A}}}{SS_{\mathrm{total}}}, \tag{14}$$

and

$$\eta^2_{\mathrm{part}} = \frac{SS_{\mathrm{A}}}{SS_{\mathrm{A}} + SS_{\mathrm{within}}} \tag{15}$$

for the nonpartial and the partial effect of the non-nested factor A, because this measure is biased in that it overestimates the true population effect. Rather, the population effect is correctly estimated with an effect size measure named *omega square:*

$$\hat{\omega}^2_{\mathrm{A}} = \frac{\hat{\sigma}^2_{\mathrm{A}}}{\hat{\sigma}^2_{\mathrm{A}} + \hat{\sigma}^2_{\mathrm{B}} + \hat{\sigma}^2_{\mathrm{within}}} \tag{16}$$

and

$$\hat{\omega}^2_{\mathrm{part}} = \frac{\hat{\sigma}^2_{\mathrm{A}}}{\hat{\sigma}^2_{\mathrm{A}} + \hat{\sigma}^2_{\mathrm{within}}} \tag{17}$$

for the estimated nonpartial and partial population effects of the non-nested factor A. The nature of the exact formula estimating the population effect on the basis of sample statistics depends on how the individual variance components are estimated in the different analysis of variance (ANOVA) models. In the mixed model, the variance components are estimated as follows (see also Table 1):

$$\hat{\sigma}^2_{\mathrm{A}} = \frac{(p-1)(MS_{\mathrm{A}} - MS_{\mathrm{B}})}{npq}, \tag{18}$$

$$\hat{\sigma}^2_{\mathrm{B}} = (MS_{\mathrm{B}} - MS_{\mathrm{within}})/n, \tag{19}$$

and

$$\hat{\sigma}^2_{\mathrm{within}} = MS_{\mathrm{within}}. \tag{20}$$

As a result, the population effect size of factor A can be estimated by the following formula for the nonpartial effect size:

$$\hat{\omega}^2_{\mathrm{mixed}} = \frac{SS_{\mathrm{A}} - (p-1)MS_{\mathrm{B}}}{SS_{\mathrm{A}} - (p-1)MS_{\mathrm{B}} + pq(MS_{\mathrm{B}} - MS_{\mathrm{within}}) + pqnMS_{\mathrm{within}}}. \tag{21}$$

Accordingly, the partial effect size for factor A can be estimated with the formula

$$\hat{\omega}^2_{\mathrm{mixed,\,partial}} = \frac{SS_{\mathrm{A}} - (p-1)MS_{\mathrm{B}}}{SS_{\mathrm{A}} - (p-1)MS_{\mathrm{B}} + pqnMS_{\mathrm{within}}}. \tag{22}$$

In the fixed model, the population effect is also estimated by Equation 16, but the estimates of the factor A and nested factor B variance components differ from the mixed model (cf. Table 1), namely:

$$\hat{\sigma}^{2\prime}_{\mathrm{A}} = \frac{(p-1)(MS_{\mathrm{A}} - MS_{\mathrm{within}})}{npq} \tag{23}$$

and

$$\hat{\sigma}^{2\prime}_{\mathrm{B}} = \frac{q-1}{qn}(MS_{\mathrm{B}} - MS_{\mathrm{within}}). \tag{24}$$

As a consequence, the nonpartial population effect size of factor A for the fixed-effects model can be estimated by the following formula:

$$\hat{\omega}^2_{\mathrm{fixed}} = \frac{SS_{\mathrm{A}} - (p-1)MS_{\mathrm{within}}}{SS_{\mathrm{total}} + MS_{\mathrm{within}}}. \tag{25}$$

The corresponding estimated partial population effect can be estimated with the following formula:

$$\hat{\omega}^2_{\mathrm{fixed,\,partial}} = \frac{SS_{\mathrm{A}} - (p-1)MS_{\mathrm{within}}}{SS_{\mathrm{A}} - (p-1)MS_{\mathrm{within}} + pqnMS_{\mathrm{within}}}. \tag{26}$$

## Statistical Power in Nested Factor Designs

It is important to consider the implication of nested factor designs on the statistical power of a study. The consequences a nested factor design will have on the power of a study will vary dramatically depending on whether the nested factor

is modeled correctly as a random or as a fixed factor. In the mixed-effects model, statistical power mainly depends on the number of levels of the nested factor, whereas power is largely independent of the number of subjects within each level of the nested factor. In fact, a mixed-model ANOVA with, for instance, a nested factor with two levels nested within the levels of a higher order factor with two levels for each treatment essentially has the statistical power of a *t* test with two degrees of freedom. In the mixed model, power is also negatively related to the magnitude of the effect of the nested factor. Studies with random nested factors should be designed accordingly with a sufficient number of levels of the nested factor, especially if a researcher expects large effects of the nested factor (i.e., a large intraclass correlation).

In the fixed model, statistical power is mainly determined by the number of subjects and remains largely unaffected by the number of levels of the nested factor. Moreover, the power of the fixed-model test increases with increasing nested factor effects because the fixed-effects model residualizes the *F*-test denominator (expected within subject variance) for nested factor variance.

## Criteria to Determine the Correct Model

Two principal possibilities exist for dealing with nested effects: Nested factors can be treated as random factors leading to a mixed-model ANOVA, or nested factors might be treated as fixed factors leading to a fixed model ANOVA. There are potential risks associated with choosing the incorrect model in nested factor designs. The incorrect use of the fixed model might lead to overestimations of effect sizes and inflated Type I error rates. In contrast, the incorrect use of the mixed model might lead to serious underestimations of effect sizes and inflated Type II errors (lack of power). It is, therefore, important to choose the correct model to analyze a nested-factor design.

If the levels of a nested factor have been randomly sampled from a universe of population levels and the goal of a researcher is to generalize to this universe of levels, the mixed model has to be used. Because the generalization of results is commonly recognized as an important aim of statistical hypothesis testing, many authors emphasize that nested factors should be treated as random effects by default. Nested factors should also be treated as random if the levels of the nested factors are randomly assigned to the levels of the nonnested factor. In the absence of random sampling from a population, random assignment can be used as a basis of statistical inference. Under the random assignment model, the statistical inference can be interpreted as applying to possible rerandomizations of the subjects in the sample.

If a researcher seeks to make an inference about the specific levels of the nested factors included in the study, a fixed-effects model should be used. Any (statistical) inference made on the basis of the fixed model is restricted to the specific levels of the nested factor as they were realized in the study.

The question of which model should be used in the absence of random sampling and random assignment is debatable. Some authors argue that the mixed model should be used regardless of whether random sampling or random assignment is involved. Other authors argue that in this case, a mixed-effects model is not justified and the fixed-effects model should be used, with an explicit acknowledgement that it does not allow a generalization of the obtained results. The choice between the mixed and the fixed model is less critical if the effects of the nested factor are zero. In this case, the mixed and the fixed model reach the same conclusions when the null hypothesis is true even if the mixed model is assumed to be a valid statistical model for the study. In particular, the mixed model does not lead to inflated Type I error levels. The fixed-effects analysis, however, can have dramatically greater power when the alternative hypothesis is true. It has to be emphasized, however, that any choice between the mixed and the fixed model should not be guided by statistical power considerations alone.

## Consequences of Ignoring Nested Factors

Although the choice between the two different models to analyze nested factor designs may be difficult, ignoring the nested factor is always a wrong decision. If the mixed model is the correct model and there are nested factor effects (i.e., the intraclass correlation is different from zero), then ignoring a nested factor, and thus the dependence of observations within the subjects within the

levels of the nested factor, leads to inflated Type I error rates and an overestimation of population effects. Some authors have suggested that after a preliminary test (with a liberal alpha level) shows that there are no significant nested factor effects, it is safe to remove the nested factor from the analysis. Monte-Carlo studies have shown, however, that these preliminary tests are typically not powerful enough (even with a liberal alpha level) to detect meaningful nested-factor effects.

However, if the fixed-effects model correctly describes the data, ignoring the nested factor will lead to an increase in Type II error levels (i.e., a loss in statistical power) and an underestimation of population effects. Both tendencies are positively related to the magnitude of the nested factor effect.

*Matthias Siemer*

***See also*** Cluster Sampling; Fixed-Effects Models; Hierarchical Linear Modeling; Intraclass Correlation; Mixed-and Random-Effects Models; Multilevel Modeling; Random-Effects Models

**Further Readings**

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data. A model comparison perspective* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Siemer, M., & Joormann, J. (2003). Power and measures of effect size in analysis of variance with fixed versus random nested factors. *Psychological Methods, 8,* 497–517.

Wampold, B. E., & Serlin, R. C. (2000). The consequence of ignoring a nested factor on measures of effect size in analysis of variance. *Psychological Methods, 5,* 425–433.

Zucker, D. M. (1990). An analysis of variance pitfall: The fixed effects analysis in a nested design. *Educational and Psychological Measurement, 50,* 731–738.

# Network Analysis

Network analysis elicits and models perceptions of the causes of a phenomenon. Typically, respondents are provided with a set of putative causal factors for a focal event and are asked to consider the relationships between these factors. These relationships are illustrated in a diagrammatic network consisting of nodes (i.e., causal factors) and arcs representing the relationships between nodes. The technique captures the complexities of people's cognitive representations of causal attributions for a given phenomenon. This entry discusses the history, techniques, applications, and limitations of network analysis.

## History

Network analysis was developed to account for individuals' relatively complex and sophisticated explanations of human behavior. It is underpinned by the notion of a *perceived causal structure,* which Harold Kelly described as being implicit in the cognitive representation of attributions. The perceived causal structure constitutes a temporally ordered network of interconnected causes and effects. Properties of the structure include the following: direction (past–future), extent (proximal–distal), patterning (simple–complex), components of varying stability–instability, and features ranging from actual to potential. The structure produced might be sparse or dense in nature, depending on the number of causal factors identified. Network analysis comprises a group of techniques developed in sociology and social anthropology, and it provides a method for generating and analyzing perceived causal networks, their structural properties, and the complex chains of relationships between causes and effects.

## Network Analysis Techniques

Network analysis can be conducted using semi-structured interviews, diagram methods, and matrix methods. Although interviews provide detailed individual networks, difficulties arise in that individual structures cannot be combined, and causal structures of different groups cannot be compared. The diagram method involves either the spatial arrangement of cards containing putative causes or the participant directly drawing the structure. Participants can both choose from a given set of potential causal factors and incorporate other personally relevant factors into their network. In addition, the strength of causal paths can be rated. Although these methods have the virtue of

ensuring only the most important causal links are elicited, they might potentially oversimplify respondents' belief structures, often revealing only sparse networks.

The matrix technique employs an adjacency grid with the causes of a focal event presented vertically and horizontally along its top and side. Participants rate the causal relationship for every pairwise combination. Early studies used a binary scale to indicate the presence/absence of causal links; however, this method does not reveal the strength of the causal links. Consequently, recent studies have used Likert scales whereby participants rate the strength of each causal relationship. A criterion is applied to these ratings to establish which of the resulting causal links should be regarded as consensually endorsed and, therefore, contributing to the network.

Early studies adopted a minimum systems criterion (MSC), the value at which all causes are included in the system, to determine the network nodes. Accordingly, causal links are added hierarchically to the network, in the order of mean strength, until the MSC is reached. It is generally accompanied by the cause-to-link ratio, which is the ratio of the number of extra links required to include a new cause in the network. Network construction stops if this requirement is too high, reducing overall endorsement of the network. An alternative criterion is inductive eliminative analysis (IEA), wherein every network produced when working toward the MSC is checked for endorsement. Originally developed to deal with binary adjacency matrices, networks were deemed consensual if endorsed by at least 50% of participants. However, the introduction of Likert scales necessitated a modified form of IEA, whereby an item average criterion (IAC) was adopted. The mean strength of a participant's endorsement of all items on a network must be above the IAC, which is usually set at 3 or 4 on a 5-point scale, depending on the overall link strength. In early research, the diagrammatic networks produced using these methods were topological, not spatial. However, recent studies have subjected the matrices of causal ratings to multidimensional scaling analysis to determine the spatial structure of networks. Thus, proximal and distal effects can be easily represented. The matrix method has the advantage of ensuring the exhaustive investigation of all possible links, and as it does not rely on participants' recall, it would be expec-ted to produce more reliable results.

## Applications

Network analysis has been applied to diverse areas to analyze belief structures. Domains that have been examined include lay understandings of social issues (e.g., loneliness, poverty), politics (e.g., the 2nd Iraq war, September 11th), and more recently illness attributions for health problems (e.g., work-based stress, coronary heart disease, lower back pain, and obesity).

The hypothetical network (Figure 1) illustrates some properties of network analysis. For example, the illness is believed to have three causes: stress, smoking, and family history. Both stress and smoking are proximal causes, whereas family history is a more distal cause. In addition to a direct effect of stress, the network shows a belief that stress also has an indirect effect on illness, as stress causes smoking. Finally, there is a reciprocal relationship (bidirectional arrow) between the illness and stress, such that stress causes the illness, and having the illness causes stress.

## Limitations

There are several unresolved issues regarding the establishment of networks and the selection of cut-off points. Comparative network analysis studies are necessary to compare and evaluate the differential effectiveness of the individual network analysis methods. The criteria for selection of cut-off points for the network, such as the MSC and cause to link, have also been criticized as atheoretical, producing extremely large networks that represent an aggregate rather than a consensual solution. Although IEA resolves some of these issues,

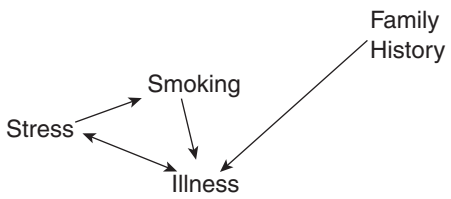



**Figure 1** Example of Network Diagram

producing networks that tend to be more consensual, smaller, and easier to interpret, the cut-off points (50% criterion and IAC) are established arbitrarily and, thus, might be contested.

*Amy Brogan and David Hevey*

***See also*** Cause and Effect; Graphical Display of Data; Likert Scaling

**Further Readings**

Kelley, H. H. (1983). Perceived causal structures. In J. Jaspars, F. D. Fincham, & M. Hewstone (Eds.), *Attribution theory and research: Conceptual, developmental and social dimensions* (pp. 343–369). London: Academic Press.

Knoke, D., & Kuklinski, J. H. (1982). *Network analysis*. Beverly Hills, CA: Sage.

# Newman–Keuls Test and Tukey Test

An analysis of variance (ANOVA) indicates whether several means come from the same population. Such a procedure is called an *omnibus* test, because it tests the whole set of means at once (*omnibus* means "for all" in Latin). In an ANOVA omnibus test, a significant result indicates that at least two groups differ from each other, but it does not identify the groups that differ. So an ANOVA is generally followed by an analysis whose goal is to identify the pattern of differences in the results. This analysis is often performed by evaluating all the pairs of means to decide which ones show a significant difference. In a general framework, this approach, which is called a pairwise comparison, is a specific case of an "a posteriori contrast analysis," but it is specific enough to be studied in itself. Two of the most common methods of pairwise comparisons are the Tukey test and the Newman–Keuls test. Both tests are based on the "Studentized range" or "Student's $q$." They differ in that the Newman–Keuls test is a sequential test designed to have more power than the Tukey test.

Choosing between the Tukey and Newman–Keuls tests is not straightforward and there is no consensus on this issue. The Newman–Keuls test is most frequently used in psychology, whereas the Tukey test is most commonly used in other disciplines. An advantage of the Tukey test is to keep the level of the Type I error (i.e., finding a difference when none exists) equal to the chosen alpha level (e.g., $\alpha = .05$ or $\alpha = .01$). An additional advantage of the Tukey test is to allow the computation of confidence intervals for the differences between the means. Although the Newman–Keuls test has more power than the Tukey test, the exact value of the probability of making a Type I error of the Newman–Keuls test cannot be computed because of the sequential nature of this test. In addition, because the criterion changes for each level of the Newman–Keuls test, confidence intervals cannot be computed around the differences between means. Therefore, selecting whether to use the Tukey or Newman–Keuls test depends on whether additional power is required to detect significant differences between means.

### *Studentized Range and Student's* q

Both the Tukey and Newman–Keuls tests use a sampling distribution derived by Willam Gosset (who was working for Guiness and decided to publish under the pseudonym of "Student" because of Guiness's confidentiality policy). This distribution, which is called the Studentized Range or Student's $q$, is similar to a $t$-distribution. It corresponds to the sampling distribution of the *largest* difference between two means coming from a set of $A$ means (when $A = 2$, the $q$ distribution corresponds to the usual Student's $t$).

In practice, one computes a criterion denoted $q_{\text{observed}}$, which evaluates the difference between the means of two groups. This criterion is computed as

$$q_{\text{observed}} = \frac{M_{i\cdot} - M_{j\cdot}}{\sqrt{MS_{\text{error}}\left(\frac{1}{S}\right)}}, \qquad (1)$$

where $M_i$ and $M_j$ are the group means being compared, $MS_{\text{error}}$ is the mean square error from the previously computed ANOVA (i.e., this is the mean square used for the denominator of the omnibus $F$ ratio), and $S$ is the number of observations per group (the groups are assumed to be of equal size).

Once the $q_{\text{observed}}$ is computed, it is then compared with a $q_{\text{critical}}$ value from a table of critical values (see Table 1). The value of $q_{\text{critical}}$ depends on the $\alpha$-level, the degrees of freedom $\nu = N - K$,

**Table 1** Table of Critical Values of the Studentized Range $q$

| | *R = Range (Number of Groups)* | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $v_2$ | *2* | *3* | *4* | *5* | *6* | *7* | *8* | *9* | *10* | *12* | *14* | *16* | *18* | *20* |
| *6* | 3.46 | 4.34 | 4.90 | 5.30 | 5.63 | 5.90 | 6.12 | 6.32 | 6.49 | 6.79 | 7.03 | 7.24 | 7.43 | 7.59 |
| | 5.24 | 6.33 | 7.03 | 7.56 | 7.97 | 8.32 | 8.61 | 8.87 | 9.10 | 9.10 | 9.48 | 10.08 | 10.32 | 10.54 |
| *7* | 3.34 | 4.16 | 4.68 | 5.06 | 5.36 | 5.61 | 5.82 | 6.00 | 6.16 | 6.43 | 6.66 | 6.85 | 7.02 | 7.17 |
| | 4.95 | 5.92 | 6.54 | 7.01 | 7.37 | 7.68 | 7.94 | 8.10 | 8.37 | 8.71 | 9.00 | 9.24 | 9.46 | 9.65 |
| *8* | 3.26 | 4.04 | 4.53 | 4.89 | 5.17 | 5.40 | 5.60 | 5.77 | 5.92 | 6.18 | 6.39 | 6.57 | 6.73 | 6.87 |
| | 4.75 | 5.64 | 6.20 | 6.62 | 6.96 | 7.24 | 7.47 | 7.68 | 7.86 | 8.18 | 8.44 | 8.66 | 8.85 | 9.03 |
| *9* | 3.20 | 3.95 | 4.41 | 4.76 | 5.02 | 5.24 | 5.43 | 5.59 | 5.74 | 5.98 | 6.19 | 6.36 | 6.51 | 6.64 |
| | 4.60 | 5.43 | 5.96 | 6.35 | 6.66 | 6.91 | 7.13 | 7.33 | 7.49 | 7.78 | 8.03 | 8.23 | 8.41 | 8.57 |
| *10* | 3.15 | 3.88 | 4.33 | 4.65 | 4.91 | 5.12 | 5.30 | 5.46 | 5.60 | 5.83 | 6.03 | 6.19 | 6.34 | 6.47 |
| | 4.48 | 5.27 | 5.77 | 6.14 | 6.43 | 6.67 | 6.87 | 7.05 | 7.21 | 7.49 | 7.71 | 7.91 | 8.08 | 8.23 |
| *11* | 3.11 | 3.82 | 4.26 | 4.57 | 4.82 | 5.03 | 5.20 | 5.35 | 5.49 | 5.71 | 5.90 | 6.06 | 6.20 | 6.33 |
| | 4.39 | 5.15 | 5.62 | 5.97 | 6.25 | 6.48 | 6.67 | 6.84 | 6.99 | 7.25 | 7.47 | 7.65 | 7.81 | 7.95 |
| *12* | 3.08 | 3.77 | 4.20 | 4.51 | 4.75 | 4.95 | 5.12 | 5.27 | 5.39 | 5.62 | 5.80 | 5.95 | 6.09 | 6.21 |
| | 4.32 | 5.05 | 5.50 | 5.84 | 6.10 | 6.32 | 6.51 | 6.67 | 6.81 | 7.06 | 7.27 | 7.44 | 7.59 | 7.73 |
| *13* | 3.06 | 3.73 | 4.15 | 4.45 | 4.69 | 4.88 | 5.05 | 5.19 | 5.32 | 5.53 | 5.71 | 5.86 | 6.00 | 6.11 |
| | 4.26 | 4.96 | 5.40 | 5.73 | 5.98 | 6.19 | 6.37 | 6.53 | 6.67 | 6.90 | 7.10 | 7.27 | 7.42 | 7.55 |
| *14* | 3.03 | 3.70 | 4.11 | 4.41 | 4.64 | 4.83 | 4.99 | 5.13 | 5.25 | 5.46 | 5.64 | 5.79 | 5.92 | 6.03 |
| | 4.21 | 4.89 | 5.32 | 5.63 | 5.88 | 6.08 | 6.26 | 6.41 | 6.54 | 6.77 | 6.96 | 7.13 | 7.27 | 7.40 |
| *15* | 3.01 | 3.67 | 4.08 | 4.37 | 4.59 | 4.78 | 4.94 | 5.08 | 5.20 | 5.40 | 5.57 | 5.72 | 5.85 | 5.96 |
| | 4.17 | 4.84 | 5.25 | 5.56 | 5.80 | 5.99 | 6.16 | 6.31 | 6.44 | 6.66 | 6.85 | 7.00 | 7.14 | 7.26 |
| *16* | 3.00 | 3.65 | 4.05 | 4.33 | 4.56 | 4.74 | 4.90 | 5.03 | 5.15 | 5.35 | 5.52 | 5.66 | 5.79 | 5.90 |
| | 4.13 | 4.79 | 5.19 | 5.49 | 5.72 | 5.92 | 6.08 | 6.22 | 6.35 | 6.56 | 6.74 | 6.90 | 7.03 | 7.15 |
| *17* | 2.98 | 3.63 | 4.02 | 4.30 | 4.52 | 4.70 | 4.86 | 4.99 | 5.11 | 5.31 | 5.47 | 5.61 | 5.73 | 5.84 |
| | 4.10 | 4.74 | 5.14 | 5.43 | 5.66 | 5.85 | 6.01 | 6.15 | 6.27 | 6.48 | 6.66 | 6.81 | 6.94 | 7.05 |
| *18* | 2.97 | 3.61 | 4.00 | 4.28 | 4.49 | 4.67 | 4.82 | 4.96 | 5.07 | 5.27 | 5.43 | 5.57 | 5.69 | 5.79 |
| | 4.07 | 4.70 | 5.09 | 5.38 | 5.60 | 5.79 | 5.94 | 6.08 | 6.20 | 6.41 | 6.58 | 6.73 | 6.85 | 6.97 |
| *19* | 2.96 | 3.59 | 3.98 | 4.25 | 4.47 | 4.65 | 4.79 | 4.92 | 5.04 | 5.23 | 5.39 | 5.53 | 5.65 | 5.75 |
| | 4.05 | 4.67 | 5.05 | 5.33 | 5.55 | 5.73 | 5.89 | 6.02 | 6.14 | 6.34 | 6.51 | 6.65 | 6.78 | 6.89 |
| *20* | 2.95 | 3.58 | 3.96 | 4.23 | 4.45 | 4.62 | 4.77 | 4.90 | 5.01 | 5.20 | 5.36 | 5.49 | 5.61 | 5.71 |
| | 4.02 | 4.64 | 5.02 | 5.29 | 5.51 | 5.69 | 5.84 | 5.97 | 6.09 | 6.29 | 6.45 | 6.59 | 6.71 | 6.82 |
| *24* | 2.92 | 3.53 | 3.90 | 4.17 | 4.37 | 4.54 | 4.68 | 4.81 | 4.92 | 5.10 | 5.25 | 5.38 | 5.44 | 5.59 |
| | 3.96 | 4.55 | 4.91 | 5.17 | 5.37 | 5.54 | 5.69 | 5.81 | 5.92 | 6.11 | 6.26 | 6.39 | 6.51 | 6.61 |
| *30* | 2.89 | 3.49 | 3.85 | 4.10 | 4.30 | 4.46 | 4.60 | 4.72 | 4.82 | 5.00 | 5.15 | 5.27 | 5.38 | 5.48 |
| | 3.89 | 4.45 | 4.80 | 5.05 | 5.24 | 5.40 | 5.54 | 5.65 | 5.76 | 5.93 | 6.08 | 6.20 | 6.31 | 6.41 |
| *40* | 2.86 | 3.44 | 3.79 | 4.04 | 4.23 | 4.39 | 4.52 | 4.63 | 4.73 | 4.90 | 5.04 | 5.16 | 5.27 | 5.36 |
| | 3.82 | 4.37 | 4.70 | 4.93 | 5.11 | 5.26 | 5.39 | 5.50 | 5.60 | 5.76 | 5.90 | 6.02 | 6.12 | 6.21 |
| *60* | 2.83 | 3.40 | 3.74 | 3.98 | 4.16 | 4.31 | 4.44 | 4.55 | 4.65 | 4.81 | 4.94 | 5.06 | 5.15 | 5.24 |
| | 3.76 | 4.28 | 4.59 | 4.82 | 4.99 | 5.13 | 5.25 | 5.36 | 5.45 | 5.60 | 5.73 | 5.84 | 5.93 | 6.02 |
| *120* | 2.80 | 3.36 | 3.68 | 3.92 | 4.10 | 4.24 | 4.36 | 4.47 | 4.56 | 4.71 | 4.84 | 4.95 | 5.04 | 5.13 |
| | 3.70 | 4.20 | 4.50 | 4.71 | 4.87 | 5.01 | 5.12 | 5.21 | 5.30 | 5.44 | 5.56 | 5.66 | 5.75 | 5.83 |
| $\infty$ | 2.77 | 3.31 | 3.63 | 3.86 | 4.03 | 4.17 | 4.29 | 4.39 | 4.47 | 4.62 | 4.74 | 4.85 | 4.93 | 5.01 |
| | 3.64 | 4.12 | 4.40 | 4.60 | 4.76 | 4.88 | 4.99 | 5.08 | 5.16 | 5.29 | 5.40 | 5.49 | 5.57 | 5.65 |

*Note:* Studentized range $q$ distribution table of critical values for $\alpha = .05$ and $\alpha = .01$.

where $N$ is the total number of participants and $K$ is the number of groups, and on a parameter $R$, which is the number of means being tested. For example, in a group of $K = 5$ means ordered from smallest to largest,

$$M_1 < M_2 < M_3 < M_4 < M_5$$

$R = 5$ when comparing $M_5$ with $M_1$; however, $R = 3$ when comparing $M_3$ with $M_1$.

### *F range*

Some statistics textbooks refer to a pseudo-$F$ distribution called the "$F$ range" or "$F_{\text{range}}$," rather than the Studentized $q$ distribution. The $F_{\text{range}}$ can be computed easily from $q$ using the following formula:

$$F_{\text{range}} = \frac{q^2}{2}. \tag{2}$$

### *Tukey Test*

For the Tukey test, $q_{\text{observed}}$ (see Equation 1) is computed between any pair of means that need to be tested. Then, $q_{\text{critical}}$ value is determined using $R =$ total number of means. The $q_{\text{critical}}$ value is the same for all pairwise comparisons. Using the previous example, $R = 5$ for all comparisons.

### *Newman–Keuls Test*

The Newman–Keuls test is similar to the Tukey test, except that the Newman–Keuls test is a sequential test in which $q_{\text{critical}}$ depends on the range of each pair of means. To facilitate the exposition, we suppose that the means are ordered from the smallest to the largest. Hence, $M_1$ is the smallest mean and $M_{\text{A}}$ is the largest mean.

The Newman–Keuls test starts exactly like the Tukey test. The largest difference between two means is selected. The range of this difference is $R = A$. A $q_{\text{observed}}$ is computed using Equation 1, and that value is compared with the critical value, $q_{\text{critical}}$, in the critical values table using $\alpha$, $\nu$, and $R$. The null hypothesis can be rejected if $q_{\text{observed}}$ is greater than $q_{\text{critical}}$. If the null hypothesis cannot be rejected, then the test stops here because not rejecting the null hypothesis for the largest difference implies not rejecting the null hypothesis for any other difference.

If the null hypothesis is rejected for the largest difference, the two differences with a range of $A - 1$ are examined. These means will be tested with $R = A - 1$. When the null hypothesis for a given pair of means cannot be rejected, none of the differences included in that difference will be tested. If the null hypothesis is rejected, then the procedure is reiterated for a range of $A - 2$ (i.e., $R = A - 2$). The procedure is reiterated until all means have been tested or have been declared nonsignificant by implication.

It takes some experience to determine which comparisons are implied by other comparisons. Figure 1 describes the structure of implication for a set of 5 means numbered from 1 (the smallest) to 5 (the largest). The pairwise comparisons implied by another comparison are obtained by following the arrows. When the null hypothesis cannot be rejected for one pairwise comparison, then all the comparisons included in it are crossed out so that they are not tested.

### *Example*

An example will help describe the use of the Tukey and Newman–Keuls tests and Figure 1. We will use the results of a (fictitious) replication of a classic experiment on eyewitness testimony by Elizabeth F. Loftus and John C. Palmer. This experiment tested the influence of question wording on the answers given by eyewitnesses. The authors presented a film of a multiple-car accident to their participants. After viewing the film, participants were asked to answer several specific questions about the accident. Among the questions, one question about the speed of the car was presented in five different versions:

1. *Hit*: About how fast were the cars going when they *hit* each other?
2. *Smash*: About how fast were the cars going when they *smashed* into each other?
3. *Collide*: About how fast were the cars going when they *collided* with each other?
4. *Bump*: About how fast were the cars going when they *bumped* into each other?
5. *Contact*: About how fast were the cars going when they *contacted* each other?

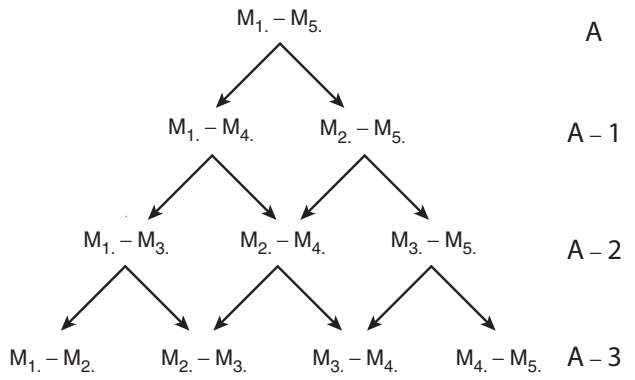



**Figure 1** Structure of Implication of the Pairwise Comparisons When $A$ = 5 for the Newman–Keuls Test

*Notes:* Means are numbered from 1 (the smallest) to 5 (the largest). The pairwise comparisons implied by another one are obtained by following the arrows. When the null hypothesis cannot be rejected for one pairwise comparison, then all the comparisons included in it can be crossed out to omit them from testing.

**Table 2** A Set of Data to Illustrate the Tukey and Newman–Keuls Tests

| | *Experimental Group* | | | | |
|---|---|---|---|---|---|
| | *Contact* | *Hit* | *Bump* | *Collide* | *Smash* |
| | 21 | 23 | 35 | 44 | 39 |
| | 20 | 30 | 35 | 40 | 44 |
| | 26 | 34 | 52 | 33 | 51 |
| | 46 | 51 | 29 | 45 | 47 |
| | 35 | 20 | 54 | 45 | 50 |
| | 13 | 38 | 32 | 30 | 45 |
| | 41 | 34 | 30 | 46 | 39 |
| | 30 | 44 | 42 | 34 | 51 |
| | 42 | 41 | 50 | 49 | 39 |
| | 26 | 35 | 21 | 44 | 55 |
| | $M_1$ | $M_2$ | $M_3$ | $M_4$ | $M_5$ |
| $M_a$ | 30.00 | 35.00 | 38.00 | 41.00 | 46.00 |

*Notes:* $S = 10$. $MS_{\text{error}} = 80.00$.

In our replication we used 50 participants (10 in each group); their responses are given in Table 2.

### *Tukey Test*

For the Tukey test, the $q_{\text{observed}}$ values are computed between every pair of means using Equation 1. For example, taking into account that the $MS_{\text{error}}$ from the previously calculated ANOVA is 80.00, the value of $q_{\text{observed}}$ for the difference between $M_1$ and $M_2$ (i.e., "contact" and "hit") is equal to

$$q_{\text{observed}} = \frac{M_1 - M_2}{\sqrt{MS_{\text{error}}\left(\frac{1}{S}\right)}} = \frac{35.00 - 30.00}{\sqrt{80.00\left(\frac{1}{10}\right)}} = \frac{5}{\sqrt{8}} = 1.77.$$

The values of $q_{\text{observed}}$ are shown in Table 3. With Tukey's approach, each $q_{\text{observed}}$ is declared significant at the $\alpha = .05$ level (or the $\alpha = .01$ level) if it is larger than the critical value obtained for this alpha level from the table with $R = 5$ and $\nu = N - K = 45$ degrees of freedom (45 is not in

**Table 3** Absolute Values of $q_{observed}$ for the Data from Table 2

| | *Experimental Group* | | | | |
|---|---|---|---|---|---|
| | $M_1$ *Contact* 30 | $M_2$ *Hit 1* 35 | $M_3$ *Bump* 38 | $M_4$ *Collide* 41 | $M_5$ *Smash* 46 |
| $M_1 = 30$ Contact | 0 | 1.77 ns | 2.83 ns | 3.89 ns | 5.66** |
| $M_2 = 35$ Hit | | 0 | 1.06 ns | 2.12 ns | 3.89 ns |
| $M_3 = 38$ Bump | | | 0 | 1.06 ns | 2.83 ns |
| $M_4 = 41$ Collide | | | | 0 | 1.77 ns |
| $M_5 = 46$ Smash | | | | | 0 |

*Notes:* For the Tukey test, $q_{\text{observed}}$ is significant at $\alpha = .05$ (or at the $\alpha = .01$ level) if $q_{\text{observed}}$ is larger than $q_{\text{critical}} = 4.04 (q_{\text{critical}} = 4.93)$. * $p < .05$. ** $p < .01$.

**Table 4** Presentation of the Results of the Tukey Test for the Data from Table 2

| | *Experimental Group* | | | | |
|---|---|---|---|---|---|
| | $M_1$ *Contact* 30 | $M_2$ *Hit 1* 35 | $M_3$ *Bump* 38 | $M_4$ *Collide* 41 | $M_5$ *Smash* 46 |
| $M_1 = 30$ Contact | 0 | 5.00 ns | 8.00 ns | 11.00 ns | 16.00** |
| $M_2 = 35$ Hit | | 0 | 3.00 ns | 6.00 ns | 11.00 ns |
| $M_3 = 38$ Bump | | | 0 | 3.00 ns | 8.00 ns |
| $M_4 = 41$ Collide | | | | 0 | 5.00 ns |
| $M_5 = 46$ Smash | | | | | 0 |

*Notes:* * $p < .05$. ** $p < .01$. The $q_{\text{observed}}$ is greater than $q_{\text{critical}(5)}$ and $H_0$ is rejected for the largest pair.

the table so 40 is used instead). The $q_{critical(5),\alpha=.05}$ is equal to 4.04 and the $q_{\text{critical}(5),\alpha=.01}$ is equal to 4.93.

When performing pairwise comparisons, it is customary to report the table of differences between means with an indication of their significance (e.g., one star meaning significant at the .05 level, and two stars meaning significant at the .01 level). This is shown in Table 4.

### *Newman–Keuls Test*

Note that for the Newman–Keuls test, the group means are ordered from the smallest to the largest. The test starts by evaluating the largest difference that corresponds to the difference between $M_1$ and $M_5$ (i.e., "contact" and "smash"). For $\alpha = .05$, $R = 5$ and $\nu = N - K = 45$ degrees of freedom, the critical value of $q$ is 4.04 (using the $\nu$ value of 40 in the table). This value is denoted as $q_{\text{critical}(5)} = 4.04$. The $q_{\text{observed}}$ is computed from Equation 1 (see also Table 3) as

$$q_{\text{observed}} = \frac{M_5 - M_1}{\sqrt{MS_{\text{error}}\left(\frac{1}{S}\right)}} = 5.66. \qquad (3)$$

Now we proceed to test the means with a range of 4, namely the differences $(M_4 - M_1)$ and $(M_5 - M_2)$. With $\alpha = .05, R = 4$ and 45 degrees of freedom, $q_{\text{critical}(4)} = 3.79$. Both differences are declared significant at the .05 level [$q_{\text{observed}(4)} = 3.89$ in both cases]. We then proceed to test the comparisons with a range of 3. The value of $q_{\text{critical}}$ is now 3.44. The differences $(M_3 - M_1)$ and $(M_5 - M_3)$, both with a $q_{\text{observed}}$ of 2.83, are declared nonsignificant. Furthermore, the difference $(M_4 - M_2)$, with a $q_{\text{observed}}$ of 2.12, is also declared nonsignificant. Hence, the null hypothesis for these differences cannot be rejected, and all comparisons implied by these differences should be crossed out. That is, we do not test any difference with a range of $A - 3$ [i.e., $(M_2 - M_1)$, $(M_3 - M_2)$, $(M_4 - M_3)$, and $(M_5 - M_4)$]. Because the comparisons with a range of 3 have already been tested and found to be nonsignificant, any comparisons with a range of 2 will consequently be declared nonsignificant as they are implied or included in the range of 3 (i.e., the test has been performed implicitly).

As for the Tukey test, the results of the Newman–Keuls tests are often presented with the values of the pairwise differences between the means and with stars indicating the significance

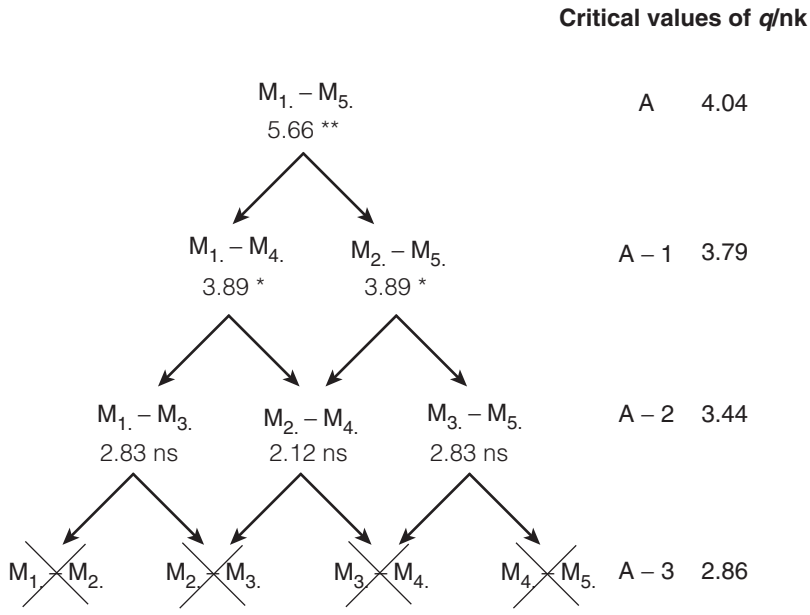



**Figure 2** Newman–Keuls Test for the Data from a Replication of Loftus & Palmer (1974)

*Note:* The number below each range is the $q_{observed}$ for that range.

**Table 5** Presentation of the Results of the Newman–Keuls Test for the Data from Table 2

| | *Experimental Group* | | | | |
|---|---|---|---|---|---|
| | $M_1$ *Contact* 30 | $M_2$ *Hit 1* 35 | $M_3$ *Bump* 38 | $M_4$ *Collide* 41 | $M_5$ *Smash* 46 |
| $M_1 = 30$ Contact | 0 | 5.00 ns | 8.00 ns | 11.00* | 16.00** |
| $M_2 = 35$ Hit | | 0 | 3.00 ns | 6.00 ns | 11.00* |
| $M_3 = 38$ Bump | | | 0 | 3.00 ns | 8.00 ns |
| $M_4 = 41$ Collide | | | | 0 | 5.00 ns |
| $M_5 = 46$ Smash | | | | | 0 |

*Notes:* * $p < .05$. ** $p < .01$.

level (see Table 5). The comparison of Table 5 and Table 4 confirms that the Newman–Keuls test is more powerful than the Tukey test.

*Hervé Abdi and Lynne J. Williams*

***See also*** Analysis of Variance (ANOVA); Bonferroni Procedure; Holm's Sequential Bonferroni Procedure; Honestly Significant Difference (HSD) Test; Multiple Comparison Tests; Pairwise Comparisons; Post Hoc Comparisons; Scheffé Test

**Further Readings**

Abdi, H., Edelman, B., Valentin, D., & Dowling, W. J. (2009). *Experimental design and analysis for psychology.* Oxford, UK: Oxford University Press.

Dudoit S., & van der Laan, M. (2008). *Multiple testing procedures with applications to genomics.* New York: Springer-Verlag.

Hochberg, Y., & Tamhane, A. C. (1987). *Multiple comparison procedures.* New York: Wiley.

Jaccard, J., Becker, M. A., & Wood, G. (1984). Pairwise multiple comparison procedures: A review. *Psychological Bulletin, 94,* 589–596.

# Nominal Scale

A nominal scale is a scale of measurement used to assign events or objects into discrete categories. This form of scale does not require the use of numeric values or categories ranked by class, but simply unique identifiers to label each distinct category. Often regarded as the most basic form of measurement, nominal scales are used to categorize and analyze data in many disciplines. Historically identified through the work of psychophysicist Stanley Stevens, use of this scale has shaped research design and continues to impact on current research practice. This entry presents key concepts, Stevens's hierarchy of measurement scales, and an example demonstrating the properties of the nominal scale.

## Key Concepts

The nominal scale, which is often referred to as the unordered categorical or discrete scale, is used to assign individual datum into categories. Categories in the nominal scale are mutually exclusive and collectively exhaustive. They are mutually exclusive because the same label is not assigned to different categories and different labels are not assigned to events or objects of the same category. Categories in the nominal scale are collectively exhaustive because they encompass the full range of possible observations so that each event or object can be categorized. The nominal scale holds two additional properties. The first property is that all categories are equal. Unlike in other scales, such as ordinal, interval, or ratio scales, categories in the nominal scale are not ranked. Each category has a unique identifier, which might or might not be numeric, which simply acts as a label to distinguish categories. The second property is that the nominal scale is invariant under any transformation or operation that preserves the relationship between individuals and their identifiers.

Some of the most common types of nominal scales used in research include sex (male/female), marital status (married or common-law/widowed/divorced/never-married), town of residence, and questions requiring binary responses (yes/no).

## Stevens's Hierarchy

In the mid-1940s, Harvard psychophysicist Stanley Stevens wrote the influential article "On

the Theory of Scales of Measurement," published in *Science* in 1946. In this article, Stevens described a hierarchy of measurement scales that includes nominal, ordinal, interval, and ratio scales. Based on basic empirical operations, mathematical group structure, and statistical procedures deemed permissible, this hierarchy has been used in textbooks worldwide and continues to shape statistical reasoning used to guide the design of statistical software packages today.

Under Stevens's hierarchy, the primary, and arguably only, use for nominal scales is to determine equality, that is, to determine whether the object of interest falls into the category of interest by possessing the properties identified for that category. Stevens argued that no other determinations were permissible, whereas others argued that even though other determinations were permissible, they would, in effect, be meaningless. A less argued property of the nominal scale is that it is invariant under any transformation. When taking attendance in a classroom, for example, those in attendance might be assigned 1, whereas those who are absent might be assigned 2. This nominal scale could be replaced by another nominal scale, where "1" is replaced by the label "present" and "2" is replaced by the label "absent." The transformation is considered invariant because the identity of each individual is preserved. Given the limited determinations deemed permissible, Stevens proposed a restriction on analysis for nominal scales. Only basic statistics are deemed permissible or meaningful for the nominal scale, including frequency, mode as the sole measure of central tendency, and contingency correlation. Despite much criticism during the past 50 years, statistical software developed during the past decade has sustained the use of Stevens's terminology and permissibility in its architecture.

## Example: Attendance in the Classroom

Again, attendance in the classroom can serve as an example to demonstrate some properties of the nominal scale. After taking attendance, the information has been recorded in the class list as illustrated in Table 1.

**Table 1** Class List for Attendance on May 1

| *Student ID* | *Arrives by School Bus* | *Attendance on May 1* |
|---|---|---|
| 001 | Yes | Absent |
| 002 | Yes | Absent |
| 003 | Yes | Present |
| 004 | Yes | Absent |
| 005 | Yes | Absent |
| 006 | Yes | Absent |
| 007 | Yes | Absent |
| 008 | Yes | Absent |
| 009 | Yes | Absent |
| 010 | No | Present |
| 011 | No | Present |
| 012 | No | Present |
| 013 | No | Present |
| 014 | No | Absent |
| 015 | No | Present |

The header row denotes the names of the variable to be categorized and each row contains an individual student record. Student 001, for example, uses the school bus and is absent on the day in question. An appropriate nominal scale to categorize class attendance would involve two categories: absent or present. Note that these categories are mutually exclusive (a student cannot be both present and absent), collectively exhaustive (the categories cover all possible observations), and each is equal in value.

Permissible statistics for the attendance variable would include frequency, mode, and contingency correlation. Using the previously provided class list, the frequency of those present is 6 and those absent is 9. The mode, or the most common observation, is "absent." Contingency tables could be constructed to answer questions about the population. If, for example, a contingency table was used

**Table 2** Contingency Table for Attendance and Arrives by School Bus

| | | *Attendance* | | |
|---|---|---|---|---|
| | | *Absent* | *Present* | *Total* |
| Arrives by school bus | Yes | 8 | 1 | 9 |
| | No | 1 | 5 | 6 |
| Total | | 9 | 6 | 15 |

to classify students using the two variables attendance and arrives by school bus, then Table 2 could be constructed.

The results of the Fisher's exact test for contingency table analysis show that those who arrive by school bus were significantly more likely to be absent than those who arrive by some other means. One might then conclude that the school bus was late.

*Deborah J. Carr*

*See also* Chi-Square Test; Frequency Table; Mode; "On the Theory of Scales of Measurement"; Ordinal Scale

**Further Readings**

Duncan, O. D. (1984). *Notes on social measurement: Historical and critical.* New York: Russell Sage Foundation.

Michell, J. (1986). Measurement scales and statistics: A clash of paradigms. *Psychological Bulletin,* 3, 398–407.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science,* 103, 677–680.

Velleman, P. F., & Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading. *The American Statistician,* 47, 65–72.

# NOMOGRAMS

Nomograms are graphical representations of equations that predict medical outcomes. Nomograms use a points-based system whereby a patient accumulates points based on levels of his or her risk factors. The cumulative points total is associated with a prediction, such as the predicted probability of treatment failure in the future. Nomograms can improve research design, and well-designed research is crucial for the creation of accurate nomograms. Nomograms are important to research design because they can help identify the characteristics of high-risk patients while highlighting which interventions are likely to have the greatest treatment effects. Nomograms have demonstrated better accuracy than both risk grouping systems and physician judgment. This improved accuracy should allow researchers to design intervention studies that have greater statistical power by targeting the enrollment of patients with the highest risk of disease. In addition, nomograms rely on well-designed studies to validate the accuracy of their predictions.

## Deriving Outcome Probabilities

All medical decisions are based on the predicted probability of different outcomes. Imagine a 35-year-old patient who presents to a physician with a 6-month history of cough. A doctor in Chicago might recommend a test for asthma, which is a common cause of chronic cough. If the same patient presented to a clinic in rural Africa, the physician might be likely to test for tuberculosis. Both physicians might be making sound recommendations based on the predicted probability of disease in their locale. These physicians are making clinical decisions based on the overall probability of disease in the population. These types of decisions are better than arbitrary treatment but treat all patients the same.

A more sophisticated method for medical decision making is *risk stratification.* Physicians will frequently assign patients to different risk groups when making treatment decisions. Risk group assignment will generally provide better predicted probabilities than estimating risk according to the overall population. In the previous cough example, a variety of other factors might impact the predicted risk of tuberculosis (e.g., fever, exposure to tuberculosis, and history of tuberculosis vaccine) that physicians are trained to explore. Most risk stratification performed in clinical practice is based on rough estimates that simply order patients into levels of risk, such as high risk, medium risk, or low risk. Nomograms provide precise probability estimates that generally make more accurate assessments of risk.

A problem with risk stratification arises when continuous variables are turned into categorical variables. Physicians frequently commit dichotomized cutoffs of continuous laboratory values to memory to guide clinical decision making. For example, blood pressure cut-offs are used to guide treatment decisions for hypertension. Imagine a new blood test called *serum marker A.* Research shows that tuberculosis patients with serum marker A levels greater than 50 are at an increased

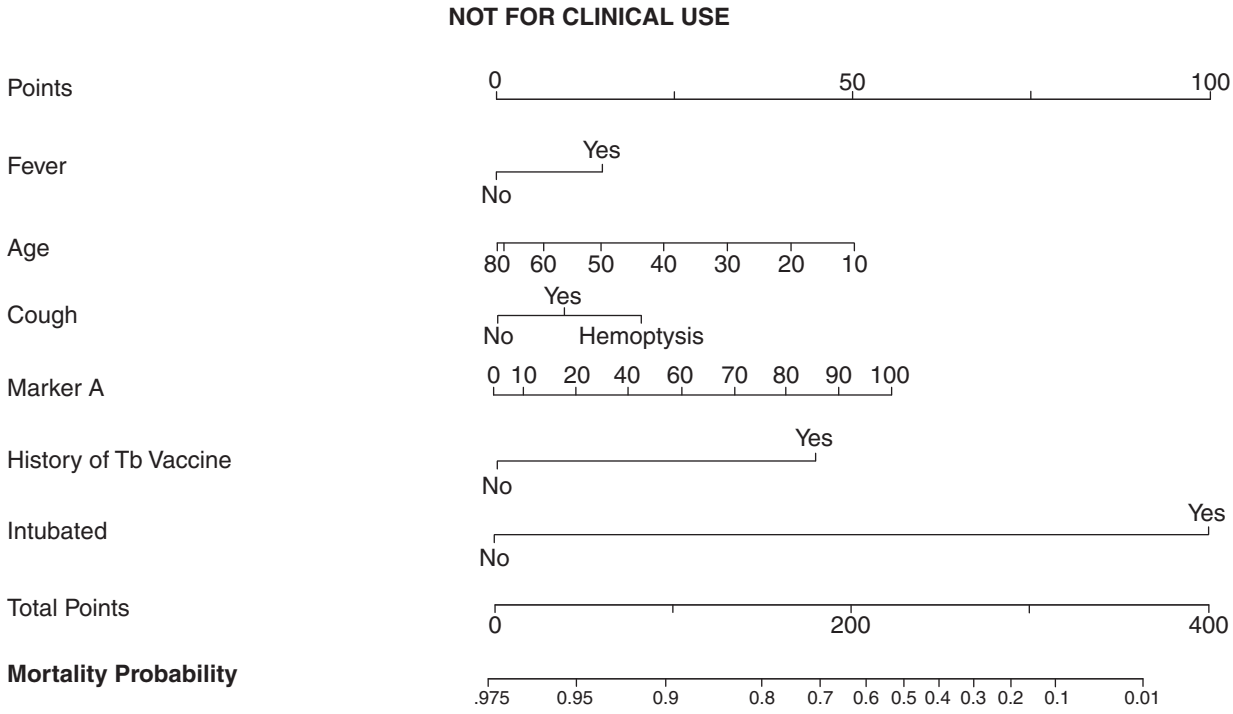



**Figure 1** Hypothetical Nomogram for Predicting Mortality Risk in Tuberculosis (Tb)

risk for dying from tuberculosis. In reality, patients with a value of 51 might have similar risks compared with patients with a value of 49. In contrast, a patient with a value of 49 would be considered to have the same low risk of a patient whose serum level of marker A is 1. Nomograms allow for predictor variables to be maintained as continuous values while allowing numerous risk factors to be considered simultaneously. In addition, more complex models can be constructed that account for interactions.

Figure 1 illustrates a hypothetical nomogram designed to predict the mortality probability for patients with tuberculosis. The directions for using the nomogram are contained in the legend. One glance at the nomogram allows the user to determine quickly which predictors have the greatest potential impact on the probability. Fever has a relatively short axis and can contribute less than 25 possible points. In contrast, whether the patient required intubation has a much greater possible impact on the predicted probability of mortality.

Nomograms like the one shown in Figure 1 are created from the coefficients obtained by the statistical model (e.g., logistic regression or Cox proportional hazards regression) and are only as precise as the paper graphics. However, the coefficients used to create the paper-based nomogram can be used to calculate the exact probability. Similarly, the coefficients can be plugged into a Microsoft Excel spreadsheet or built into a computer interface that will automatically calculate the probability based on the user inputs.

## Why Are Nomograms Important to Research Design?

Nomograms provide an improved ability to identify the correct patient population for clinical

studies. The statistical power in prospective studies with dichotomous clinical outcomes is derived from the number of events. (Enrolling excessive numbers of patients who do not develop the event of interest is an inefficient use of resources.) For instance, let us suppose that a new controversial medication has been developed for treating patients with tuberculosis. The medication shows promise in animal studies, but it also seems to carry a high risk of serious toxicity and even death in some individuals. Researchers might want to determine whether the medication improves survival in some patients with tuberculosis. The nomogram in Figure 1 could be used to identify which patients are at highest risk of dying, are most likely to benefit from the new drug, and therefore should be enrolled in a drug trial. The nomogram could also be tested in this fashion using a randomized clinical trial design. One arm of the study could be randomized to usual care, whereas the treatment arm is randomized to use the Tb nomogram and then to receive the usual care if the risk of mortality is low or the experimental drug if the risk of mortality is high.

### *Validation*

The estimated probability obtained from nomograms like the one in Figure 1 is generally much more accurate than rough probabilities obtained by risk stratification and should help both patients and physicians make better treatment decisions. However, nomograms are only as good as the data that were used for their creation. But, predicted probabilities can be graded (validated) on their ability to discriminate between pairs of patients who have different outcomes (discordant pairs). The grading can be performed using either a validation data set that was created with the same database used to create the prediction model (internal validation) or with external data (external validation). Ideally, a nomogram should be validated in an external database before it is widely used in heterogeneous patient populations.

A validation data set using the original data can be created either with the use of bootstrapping or by dividing the data set into random partitions. In the bootstrap method, a random patient is selected and a copy of the patient's data is added to the validation data set. The patient's record is maintained in the original data set and is available for subsequent random selection. The random selection of patients is continued until a dataset that is the same size as the original data has been formed. The model is applied (i.e., fit) to the bootstrap data and the model is graded on its ability to predict accurately the outcome of patients in either the original data (apparent accuracy) or the bootstrap sample (unbiased accuracy). Alternatively, the original data can be partitioned randomly. The model is fit to only a portion of the original data and the outcome is predicted in the remaining subset. The bootstrap method has the added benefit that the sample size used for the model fitting is not reduced.

### *Evaluating Model Accuracy*

As mentioned previously, the models' predictions are evaluated on their ability to discriminate between pairs of discordant patients (patients who had different outcomes). The resultant evaluation is called a concordance index or c-statistic. The concordance index is simply the proportion of the time that the model accurately assigns a higher risk to the patient with the outcome. The c-statistic can vary from 0.50 (equivalent to the flip of a coin) to 1.0 (perfect discrimination). The c-statistic provides an objective method for evaluating model accuracy, but the minimum c-statistic needed to claim that a model has good accuracy depends on the specific condition and is somewhat subjective. However, models are generally not evaluated in isolation. Models can be compared head-to-head either with one another or with physician judgment. In this case, the most accurate model can generally be identified as the one with the highest concordance index.

However, to grade a model fully, it is also necessary to determine a model's calibration. Calibration is a measure of how close a model's prediction compares with the actual outcome and is frequently displayed by plotting the predicted probability (or value) versus the actual proportion with the outcome (or actual value). The concordance index is simply a "rank" test that orders patients according to risk. A model can theoretically have a great concordance index but poor calibration. For instance, a model might rank patients

appropriately while significantly overestimating or underestimating the probability (or value) in all of the patients.

### Conclusion

Designing efficient clinical research, especially when designing prospective studies, relies on accurate predictions of the possible outcomes. Nomograms provide an opportunity for researchers to easily identify the target population that will be predicted to have the highest incidence of events and will therefore keep the necessary sample size low. Paper-based nomograms provide an excellent medium for easily displaying risk probabilities and do not require a computer or calculator. The coefficients used to construct the nomogram can be used to create a computer-based prediction tool.

However, nomograms are only as good as the data that were used in their creation, and no nomogram can provide a perfect prediction. Ultimately, the best evaluation of a nomogram is made by validating the prediction accuracy of a nomogram on an external data set and comparing the concordance index with another prediction method that was validated using the same data. The validation of nomograms provides another opportunity for research design. Prospective studies that collect all the predictor variables needed to calculate a specific nomogram are ideal for determining a nomogram's accuracy. In addition, more randomized controlled trials that compare nomogram-derived treatment recommendations versus standard of care are needed to promote the use of nomograms in medicine.

*Brian J. Wells and Michael Kattan*

***See also*** Decision Rule; Evidence-Based Decision Making; Probability, Laws of

#### Further Readings

Harrell, F. E., Jr. (1996). Multivariate prognostic models: Issues in developing models, evaluating assumptions and accuracy, and measuring and predicting errors. *Statistics in Medicine,* 15, 361.

Harrell, F. E., Jr., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982). Evaluating the yield of medical tests. *Journal of the American Medical Association,* 247, 2543–2546.

Kattan, M. W. (2003). Nomograms are superior to staging and risk grouping systems for identifying high-risk patients: Preoperative application in prostate cancer. *Current Opinion in Urology,* 13, 111–116.

## Nonclassical Experimenter Effects

Experimenter effects denominate effects where an outcome seems to be a result of an experimental intervention but is actually caused by conscious or unconscious effects the experimenter has on how data are produced or processed. This could be through inadvertently measuring one group differently from another one, treating a group of people or animals that are known to receive or to have received the intervention differently compared with the control group, or biasing the data otherwise. Normally, such processes happen inadvertently because of expectation and because participants sense the desired outcome in some way and hence comply or try to please the experimenter. Control procedures, such as blinding (keeping participants and/or experimenters unaware of a study's critical aspects), are designed to keep such effects at bay. Whenever the channels by which such effects are transmitted are potentially known or knowable, the effect is known as a *classical* experimenter effect. They normally operate through the known senses and very often by subliminal perception. If an experiment is designed to exclude such classical channels of information transfer, because it is testing some claims of anomalous cognition, and such differential effects of experimenters still happen, then these effects are called *nonclassical* experimenter effects, because there is no currently accepted model to understand how such effects might have occurred in the first place.

### Empirical Evidence

This effect has been known in parapsychological research for awhile. Several studies reported that parapsychological effects were found in some studies, whereas in other studies with the same experimental procedure, the effects were not shown. A

well-known experiment that has shown such a nonclassical experimenter effect is one where a parapsychological researcher who had previously produced replicable results with a certain experimental setup invited a skeptical colleague into her laboratory to replicate the experiment with her. They ran the same experiment together; half of the subjects were introduced to the experimental procedures by the enthusiastic experimenter and half by the skeptical experimenter. The experimental task was to influence a participant's arousal remotely, measured by electrodermal activity, via intention only according to a random sequence. The two participants were separated from each other and housed in shielded chambers. Otherwise, all procedures were the same. Although the enthusiastic researcher could replicate the previous results, the skeptical researcher produced null results. This finding occurred even though there was no way of transferring the information in the experiment itself. This result was replicated in another study in the skeptical researcher's laboratory, where again the enthusiastic researcher could replicate the findings but the skeptic could not. There are also several studies reported where more than one experimenter interacted with the participants. If these studies are evaluated separately for each experimenter, it could be shown that some experimenters find consistently significant results whereas others do not. These are not only exploratory findings because some of these studies could be repeated and the experimenter effects were hypothesized.

Another experimental example are the so-called memory-of-water effects, where Jacques Benveniste, who was a French immunologist, had claimed that water mixed with an immunogenic substance and successively diluted in steps to a point where no original molecules were present would still have a measurable effect. Blinded experiments produced some results, sometimes replicable and sometimes not. Later, he claimed that such effects can also be digitized, recorded, and played back via a digital medium. A definitive investigation could show that these effects only happened when one particular experimenter was present who was known to be indebted to Benveniste and wanted the experiments to work. Although a large group of observers with specialists from different disciplines were present, there was no indication how the individual in question could have potentially biased this blinded system, although such tampering, and hence a classical experimenter effect, could not be excluded.

The nonclassical experimenter effect has been shown repeatedly in parapsychological research. The source of this effect is unclear. If the idea behind parapsychology that intention can affect physical systems without direct interaction is at all sensible and worth any consideration, then there is no reason why the intention of an experimenter should be left out of an experimental system in question. Furthermore, one could argue that if the intention of an experimental participant could affect a system without direct interaction, then the intention of the experimenter could do the same. Strictly speaking, any nonclassical experimenter effect defies experimental control and calls into question the concept of experimental control.

## Theoretical Considerations

When it comes to understanding such effects, they are probably among the strongest empirical facts that point to a partially constructivist view of the world that is also embedded in some spiritual worldviews such as in the Buddhist, Vedanta, or other mystical concepts. Here, our mental constructs, intentions, thoughts, and wishes are not only reflections of the world or idle mental operations that might affect the world indirectly by being responsible for our future actions but also could be viewed as constituents and creators of reality itself. This is difficult to understand within the accepted scientific framework of the world. Hence, such effects and a constructivist concept of reality also point to the limits of the validity of our current worldview. For such effects to be scientifically viable concepts, researchers need to envisage a world in which mental and physical acts can interact with each other directly. Such effects make us aware of the fact that we constantly partition our world into compartments and pieces that are useful for certain purposes, for instance, for the purpose of technical control, but do not necessarily describe reality as such. In this sense, they remind us of the constructivist basis of science and the whole scientific enterprise.

*Harald Walach and Stefan Schmidt*

*See also* Experimenter Expectancy Effect; Hawthorne Effect; Rosenthal Effect

**Further Readings**

Collins, H. M. (1985). *Changing order.* Beverly Hills, CA: Sage.

Jonas, W. B., Ives, J. A., Rollwagen, F., Denman, D. W., Hintz, K., Hammer, M., et al. (2006). Can specific biological signals be digitized? *FASEB Journal, 20,* 23–28.

Kennedy, J. E., & Taddonio, J. L. (1976). Experimenter effects in parapsychological research. *Journal of Parapsychology, 40,* 1–33.

Palmer, J. (1997). The challenge of experimenter psi. *European Journal of Parapsychology, 13,* 110–125.

Smith, M. D. (2003). The role of the experimenter in parapsychological research. *Journal of Consciousness Studies, 10,* 69–84.

Walach, H., & Schmidt, S. (1997). Empirical evidence for a non-classical experimenter effect: An experimental, double-blind investigation of unconventional information transfer. *Journal of Scientific Exploration, 11,* 59–68.

Watt, C. A., & Ramakers, P. (2003). Experimenter effects with a remote facilitation of attention focusing task: A study with multiple believer and disbeliever experimenters. *Journal of Parapsychology, 67,* 99–116.

Wiseman, R., & Schlitz, M. (1997). Experimenter effects and the remote detection of staring. *Journal of Parapsychology, 61,* 197–208.

## Nondirectional Hypotheses

A nondirectional hypothesis is a type of alternative hypothesis used in statistical significance testing. For a research question, two rival hypotheses are formed. The *null hypothesis* states that there is no difference between the variables being compared or that any difference that does exist can be explained by chance. The *alternative hypothesis* states that an observed difference is likely to be genuine and not likely to have occurred by chance alone. Sometimes called a two-tailed test, a test of a *nondirectional alternative hypothesis* does not state the direction of the difference, it indicates only that a difference exists. In contrast, a directional alternative hypothesis specifies the direction of the tested relationship, stating that one variable is predicted to be larger or smaller than null value, but not both. Choosing a nondirectional or directional alternative hypothesis is a basic step in conducting a significance test and should be based on the research question and prior study in the area. The designation of a study's hypotheses should be made prior to analysis of data and should not change once analysis has been implemented.

For example, in a study examining the effectiveness of a learning strategies intervention, a treatment group and a control group of students are compared. The null hypothesis states that there is no difference in mean scores between the two groups. The nondirectional alternative hypothesis states that there is a difference between the mean scores of two groups but does not specify which group is expected to be larger or smaller. In contrast, a directional alternative hypothesis might state that the mean of the treatment group will be larger than the mean of the control group. The null and the nondirectional alternative hypothesis could be stated as follows:

*Null Hypothesis:* $H_0 : \mu 1 - \mu 2 = 0.$

*Nondirectional Alternative Hypothesis:* $H_1 : \mu 1 - \mu 2 \neq 0.$

A common application of nondirectional hypothesis testing involves conducting a $t$ test and comparing the means of two groups. After calculating the $t$ statistic, one can determine the critical value of $t$ that designates the null hypothesis rejection region for a nondirectional or two-tailed test of significance. This critical value will depend on the degrees of freedom in the sample and the desired probability level, which is usually .05. The rejection region will be represented on both sides of the probability curve because a nondirectional hypothesis is sensitive to a larger or smaller effect.

Figure 1 shows a distribution in which at the 95% confidence level, the solid regions at the top and bottom of the distribution represent 2.5% accumulated probability in each tail. If the calculated value for $t$ exceeds the critical value at either tail of the distribution, than the null hypothesis can be rejected.

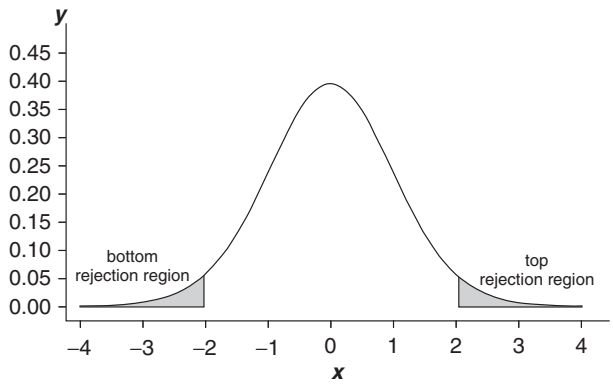



**Figure 1** Nondirectional *t* Test with *df* = 30

*Note:* Alpha = 0.05, critical value = 2.0423.

In contrast, the rejection region of a directional alternative hypothesis, or one-tailed test, would be represented on only one side of the distribution, because the hypothesis would choose a smaller or larger effect, but not both. In this instance, the critical value for *t* will be smaller because all 5% probability will be represented on one tail.

Much debate has occurred over the appropriate use of nondirectional and directional hypothesis in significance testing. Because the critical rejection values for a nondirectional test are higher than a directional test, it is a more conservative approach and the most commonly used. However, when prior research supports the use of a directional hypothesis test, a significant effect is easier to find and, thus, represents more statistical power. The researcher should state clearly the chosen alternative hypothesis and the rationale for this decision.

*Gail Tiemann, Neal Kingston, Jie Chen, and Fei Gu*

*See also* Alternative Hypotheses; Directional Hypothesis; Hypothesis

### Further Readings

Marks, M. R. (1951). Two kinds of experiment distinguished in terms of statistical operations. *Psychological Review,* 60, 179–184.

Nolan, S., & Heinzen, T. (2008). *Statistics for the behavioral sciences* (1st ed.). New York: Worth Publishers.

Pillemer, D. B. (1991). One-versus two-tailed hypothesis tests in contemporary educational research. *Educational Researcher, 20,* 13–17.

## Nonexperimental Designs

Nonexperimental designs include research designs in which an experimenter simply either describes a group or examines relationships between preexisting groups. The members of the groups are not randomly assigned and an independent variable is not manipulated by the experimenter, thus, no conclusions about causal relationships between variables in the study can be drawn. Generally, little attempt is made to control for threats to internal validity in nonexperimental designs. Nonexperimental designs are used simply to answer questions about groups or about whether group differences exist. The conclusions drawn from nonexperimental research are primarily descriptive in nature. Any attempts to draw conclusions about causal relationships based on nonexperimental research are done so post hoc.

This entry begins by detailing the differences between nonexperimental and other research designs. Next, this entry discusses types of nonexperimental designs and the potential threats to internal validity that nonexperimental designs present. Last, this entry examines the benefits of using nonexperimental designs.

### Differences Among Experimental, Quasi-Experimental, and Nonexperimental Designs

The crucial differences between the three main categories of research design lie in the assignment of participants to groups and in the manipulation of an independent variable. In experimental designs, members are randomly assigned to groups and the experimenter manipulates the values of the independent variable so that causal relationships might be established or denied. In quasi-experimental and nonexperimental designs, the groups already exist. The experimenter cannot randomly assign the participants to groups because either the groups were already established before the experimenter began his or her research or the groups are being established by someone other than the researcher for

a purpose other than the experiment. In quasi-experimental designs, the experimenter can still manipulate the value of the independent variable, even though the groups to be compared are already established. In nonexperimental designs, the groups already exist and the experimenter cannot or does not attempt to manipulate an independent variable. The experimenter is simply comparing the existing groups based on a variable that the researcher did not manipulate. The researcher simply compares what is already established. Because he or she cannot manipulate the independent variable, it is impossible to establish a causal relationship between the variables measured in a nonexperimental design.

A nonexperimental design might be used when an experimenter would like to know about the relationship between two variables, like the frequency of doctor visits for people who are obese compared with those who are of healthy weight or are underweight. Clearly, from both an ethical and logistical standpoint, an experimenter could not simply select three groups of people randomly from a population and make one of the groups obese, one of the groups healthy weight, and one of the groups underweight. The experimenter could, however, find obese, healthy weight, and underweight people and record the number of doctor visits the members of each of these groups have to look at the relationship between the variables of interest. This nonexperimental design might yield important conclusions even though a causal relationship could not clearly be established between the variables.

## Types of Nonexperimental Designs

Although the researcher does not assign participants to groups in nonexperimental design, he or she can usually still determine what is measured and when it will be measured. So despite the lack of control in aspects of the experiment that are generally important to researchers, there are still ways in which the experimenter can control the data collection process to obtain interesting and useful data. Various authors classify nonexperimental designs in a variety of ways. In the subsequent section, six types of frequently used nonexperimental designs are discussed: comparative designs, causal-comparative designs (which are also referred to as differential or *ex post facto* designs), correlational designs, developmental designs, one-group pretest–posttest designs, and finally posttest only nonequivalent group designs.

### *Comparative Designs*

In these designs, two or more groups are compared on one or more measures. The experimenter might collect quantitative data and look for statistically significant differences between groups, or the experimenter might collect qualitative data and compare the groups in a more descriptive manner. Of course, the experimenter might also use mixed methods and do both of the previously mentioned strategies. Conclusions can be drawn about whether differences exist between groups, but the reasons for the differences cannot be drawn conclusively. The study described previously regarding obese, healthy weight, and underweight people's doctor visits is an example of a comparative design.

### *Causal-Comparative, Differential, or Ex Post Facto Research Designs*

Nonexperimental research that is conducted when values of a dependent variable are compared based on a categorical independent variable is often referred to as a causal-comparative or a differential design. In these designs, the groups are determined by their values on some preexisting categorical variable, like gender. This design is also sometimes called ex post facto for that reason; the group membership is determined after the fact. After determining group membership, the groups are compared on the other measured dependent variable. The researcher then tests for statistically significant differences in the dependent variable between groups. Even though this design is referred to as causal comparative, a causal relationship cannot be established using this design.

### *Correlational Designs*

In correlational designs, the experimenter measures two or more nonmanipulated variables for each participant to ascertain whether linear relationships exist between the variables. The researcher might use the correlations to conduct

subsequent regression analyses for predicting the values of one variable from another. No conclusions about causal relationships can be drawn from correlational designs. It is important to note, however, that correlational analyses might also be used to analyze data from experimental or quasi-experimental designs.

### *Developmental Designs*

When a researcher is interested in developmental changes that occur over time, he or she might choose to examine the relationship between age and other dependent variables of interest. Clearly, the researcher cannot manipulate age, so developmental studies are often conducted using nonexperimental designs. The researcher might find groups of people at different developmental stages or ages and compare them on some characteristics. This is essentially a form of a differential or causal-comparative design in that group membership is determined by one's value of a categorical variable. Although age is not inherently a categorical variable, when people are grouped together based on categories of ages, age acts as a categorical variable.

Alternatively, the researcher might investigate one group of people over time in a longitudinal study to examine the relationship between age and the variables of interest. For example, the researcher might be interested in looking at how self-efficacy in mathematics changes as children grow up. He or she might measure the math self-efficacy of a group of students in 1st grade and then measure that same group again in the 3rd, 5th, 7th, 9th, and 11th grades. In this case, children were not randomly assigned to groups and the independent variable (age) was not manipulated by the experimenter. These two characteristics of the study qualify it as nonexperimental research.

### *One-Group Pretest–Posttest Design*

In this within-subjects design, each individual in a group is measured once before and once after a treatment. In this design, the researcher is not examining differences between groups but examining differences across time in one group. The researcher does not control for possible extraneous variables possibly causing change over time. As with all nonexperimental designs, the researcher does not control the independent variable. However, this design is generally used when a researcher knows that an intervention of some kind will be taking place in the future. Thus, although the researcher is not manipulating an independent variable, someone else is. When the researcher knows this will occur before it happens, he or she can collect pretest data, which are simply data collected before the intervention. An example of this design would be if a professor wants to study the impact of a new campus-wide recycling program that will be implemented soon. The professor might want to collect data on the amount of recycling that occurs on campus before the program and on attitudes about recycling before the implementation of the program. Then, perhaps 6 months after the implementation, the professor might want to collect the same kind of data again. Although the professor did not manipulate the independent variable of the recycling program and did not randomly assign students to be exposed to the program, conclusions about changes that occurred after the program can still be drawn. Given the lack of manipulation of the independent variable and the lack of random assignment of participants, the study is nonexperimental research.

### *Posttest-Only Nonequivalent Control Group Design*

In this type of between-subjects design, two nonequivalent groups of participants are compared. In nonexperimental research, the groups are almost always nonequivalent because the participants are not randomly assigned to groups. Because the researcher also does not control the intervention, this design is used when a researcher wants to study the impact of an intervention that already occurred. Given that the researcher cannot collect pretest data, he or she collects posttest data. However, to draw any conclusions about the posttest data, the researcher collects data from two groups, one that received the treatment or intervention, and one that did not. For example, if one is interested in knowing how participating in extracurricular sports during high school affects students' attitudes about the importance of physical fitness in adulthood, an experimenter might

survey students during the final semester of their senior year. The researcher could survey a group that participated in sports and a group that did not. Clearly, he or she could not randomly assign students to participate or not participate. In this case, he or she also could not compare the attitudes prior to participating with those after participating. Obviously with no pretest data and with groups that are nonequivalent, the conclusions drawn from these studies might be lacking in internal validity.

## Threats to Internal Validity

Internal validity is important in experimental research designs. It allows one to draw unambiguous conclusions about the relationship between two variables. When there is more than one possible explanation for the relationship between variables, the internal validity of the study is threatened. Because the experimenter has little control over potential confounding variables in nonexperimental research, the internal validity can be threatened in numerous ways.

### *Self-Selection*

The most predominant threat with nonexperimental designs is caused by the self-selection that often occurs by the participants. Participants in nonexperimental designs often join the groups to be compared because of an interest in the group or because of life circumstances that place them in those groups. For example, if a researcher wanted to compare the job satisfaction levels of people in three different kinds of careers like business, academia, and food service, he or she would have to use three groups of people that either intentionally chose those careers or ended up in their careers because of life circumstances. Either way, the employees in those careers are likely to be in those different careers because they are different in other ways, like educational background, skills, and interests. Thus, if the researcher finds differences in job satisfaction levels, they might be because the participants are in different careers or they might be because people who are more satisfied with themselves overall choose careers in business, whereas those who do not consider their satisfaction in life choose careers in academia, and those that are between careers of their choosing opt for jobs in food service. Another possibility is that people with more education are more satisfied with their jobs, and people in academia tend to be the most educated, followed by those in business and then those in food service. Thus, if the researcher found that academics are the most satisfied, it might be because of their jobs, or it might be because of their education. These proposed rationales are purely speculative; however, they demonstrate how internal validity might be threatened by self-selection. In both cases, a third variable exists that contributes to the differences between groups. Third variables can threaten internal validity in numerous ways with nonexperimental research.

### *Assignment Bias*

Like self-selection, the assignment of participants to groups in a nonrandom method can create a threat to internal validity. Although participants do not always self-select into groups used in nonexperimental designs, when they do not self-select, they are generally assigned to a group for a particular reason by someone other than the researcher. For example, if a researcher wanted to compare the vocabulary acquisition of students exposed to bilingual teachers in elementary schools, he or she might compare students taught by bilingual teachers with students taught by monolingual teachers in one school. Students might have been assigned to their classes for reasons related to their skill level in vocabulary related tasks, like reading. Thus, any relationship the researcher finds might be caused not by the exposure to a bilingual teacher but by a third variable like reading level.

### *History and Maturation*

In nonexperimental designs, an experimenter might simply look for changes across time in a group. Because the experimenter does not control the manipulation of the independent variable or group assignment, both history and maturation can affect the measures collected from the participants. Some uncontrolled event (history) might occur that might confuse the conclusions drawn by the experimenter. For example, in the job satisfaction study above, if the researcher was

looking at changes in job satisfaction over time and during the course of the study the stock market crashed, then many of those with careers in business might have become more dissatisfied with their jobs because of that event. However, a stock market crash might not have affected academics and food service workers to the same extent that it affected business workers. Thus, the conclusions that might be formed about the dissatisfaction of business employees would not have internal validity.

Similarly, in the vocabulary achievement example above, one would expect elementary students' vocabularies to improve simply because of maturation over the course of a school year. Thus, if the experimenter only examined differences in vocabulary levels for students with bilingual teachers over the course of the school year, then he or she might draw erroneous conclusions about the relationship between vocabulary performance and exposure to a bilingual teacher when in fact no relationship exists. This maturation of the students would be a threat to the internal validity of that study.

## Benefits of Using Nonexperimental Designs

Nonexperimental designs are often relied on when a researcher has a question that requires a large group that cannot easily be assigned to groups. They might also be used when the population of interest is small and hard to access. Sometimes, nonexperimental designs are used when a researcher simply wants to know something about a population but does not actually have a research hypothesis.

Although experimental designs are often used in both the hard and social sciences, there are numerous occasions in social sciences in which it is simply not possible to use an experimental design. This is especially true in fields like education or program evaluation, where the programs to be studied cannot simply be offered to a random set of participants from the population of interest. Rather, the educational program might already be in use in certain schools or classrooms. The researcher might have to determine how conclusions can be drawn based on the already established samples that are either participating or not participating in the intervention that is not controlled by the researcher. In such cases, it is preferable to use a nonexperimental design to acquire as much information about the program's effectiveness as possible rather than simply to not attempt to study the effectiveness of the program.

Even though nonexperimental designs give the experimenter little control over the experimental process, the experimenter can improve the reliability of the findings by replicating the study. Additionally, one important feature of nonexperimental designs is the possibility of stronger ecological validity than one might obtain with a controlled, experimental design. Given that nonexperimental designs are often conducted with preexisting interventions with "real people" in the "real world," rather than participants in a laboratory, the findings are often more likely to be true to other real-world situations.

*Jill H. Lohmeier*

***See also*** Experimental Design; Internal Validity; Quasi-Experimental Design; Random Assignment; Research Design Principles; Threats to Validity; Validity of Research Conclusions

### Further Readings

DePoy, E., & Gilson, S. F. (2003). *Evaluation practice.* Pacific Grove, CA: Brooks/Cole.

Gravetter, F. J., & Forzano, L. B. (2009). *Research methods for the behavioral sciences,* Belmont, CA: Wadsworth Cengage Learning.

Johnson, B., & Christensen, L. B. (2007). *Educational research*. Thousand Oaks, CA: Sage.

Kerlinger, F. N., & Lee, H. B. (2000). *Foundations of behavioral research* (4th ed.). Belmont, CA: Wadsworth Cengage Learning.

McMillan, J. H., & Schumacher, S. (2006). *Research in education: Evidence-based inquiry* (6th ed.). Boston: Pearson Education, Inc.

Payne, D. A. (1994). *Designing educational project and program evaluations: A practical overview based on research and experience*. Dordrecht, the Netherlands: Kluwer Academic Publishers.

Punch, K. F. (1998) *Introduction to social research: Quantitative & qualitative approaches*. Thousand Oaks, CA: Sage.

Smith, E. R., & Mackie, D. M. (2007). *Social psychology* (3rd ed.). Philadelphia: Psychology Press.

Spector, P. E. (1990). *Research designs, series: Quantitative applications in the social sciences.* Newbury Park, CA: Sage.

# Nonparametric Statistics

Nonparametric statistics refer to methods of measurement that do not rely on assumptions that the data are drawn from a specific distribution. Nonparametric statistical methods have been widely used in various kinds of research designs to make statistical inferences. In practice, when the normality assumption on the measurements is not satisfied, parametric statistical methods might provide misleading results. In contrast, nonparametric methods make much less stringent distributional assumptions on the measurements. They are valid methods regardless of the underlying distributions of the observations. Because of this attractive advantage, ever since the first introduction of nonparametric tests in the last century, many different types of nonparametric tests have been developed to analyze various types of experimental designs. Such designs encompass one-sample design, two-sample design, randomized-block design, two-way factorial design, repeated measurements design, and high-way layout. The observations in each experimental condition could be equal or unequal. The targeted inferences include the comparisons of treatment effects, the existence of interaction effects, ordered inferences of the effects, and multiple comparisons of the effects. All these methods share the same feature that instead of using the actual observed measurements, they used the ranked values to form the statistics. By discarding the actual measurements, the methods gain the robustness to the underlying distributions and the potential contamination of outliers. This gain of robustness is only at the price of losing a relatively small amount of efficiency. In this entry, a brief review of the existing nonparametric methods is provided to facilitate the application of them in practical settings.

## Tests for One or Multiple Populations

The Mann–Whitney–Wilcoxon (MWW) test is a nonparametric test to determine whether two samples of observations are drawn from the same distribution. The null hypothesis specifies that the two probability distributions are identical. It is one of the best-known nonparametric tests. It was first proposed by Frank Wilcoxon in 1945 for equal sample sizes, and it was later extended to arbitrary sample sizes by Henry B. Mann and Donald R. Whitney in 1947. To obtain the statistic, the observations are first ranked without regard to which sample they are in. Then for samples 1 and 2, the sum of ranks $R_1$ and $R_2$, respectively, are computed. The statistic takes the form of

$$U = \mathrm{MIN}[R_1 - N_1(N_1-1)/2, R_2 - N_2(N_2-1)/2],$$

where $N_1, N_2$ denotes the samples sizes. For small samples, the distribution of the statistic is tabulated. However, for sample sizes greater than 20, the statistic can be normalized into

$$z = \frac{U - \dfrac{N_1 N_2}{2}}{\sqrt{\dfrac{N_1 N_2 (N_1 + N_2 + 1)}{12}}}.$$

The significance of the normalized statistic $z$ can be assessed using the standard normal table. As a test to compare two populations, MWW is in spirit very similar to the parametric two-sample $t$ test. In comparison, the parametric $t$ test is more powerful if the data are drawn from normal distribution. In contrast, if the distributional assumption is violated in practice, then MWW is more powerful than its parametric counterpart. In terms of efficiency, under the normality assumption, the efficiency of MWW test is 95% of that of $t$ test, which implies that to achieve the same power, the $t$ test will need 5% less data points than the MWW test. Under other non-normal, especially heavy-tailed distributions, the efficiency of MWW could be much higher.

For the case of two related samples or repeated measurements on a single sample, the Wilcoxon signed-rank test is a nonparametric alternative to the paired Student's $t$ test. The null hypothesis to be tested is that the median of the paired differences is equal to zero. Let $(X_1, Y_1), \ldots, (X_N, Y_N)$ denote the paired observations. The researcher computes the differences $d_i = X_i - Y_i$, $i = 1, \ldots, N$, and omits those differences with zero values. Then, the remaining differences are ranked without regard to sign. The researcher then computes $W_+$ and $W_-$ as

the sums of the ranks corresponding to the positive and negative differences. If the alternative hypothesis specifies the median is greater, less than, or unequal, then the test statistic will be $W_+, W_-$, or $\text{MIN}(W_-, W_+)$, respectively. For small samples with $N$ less than 30, the table of critical values are tabulated, whereas $N$ is large, normal approximation could be used to assess the significance. Note that the Wilcoxon signed-rank test can be used directly on one sample to test whether the population median is zero or not.

For comparisons of multiple populations, the nonparametric counterpart of the analysis of variance (ANOVA) test is the Kruskal–Wallis $k$-sample test proposed by William H. Kruskal and W. Allen Wallis in 1952. Given independent random samples of sizes $N_1, N_2, \ldots, N_K$, drawn from $k$ populations, the null hypothesis is that all the $k$ populations are identical and have the same median; the alternative hypothesis is that at least one of the populations has a median different from the others. Let $N$ denote the total number of measurements in the $k$ samples, $N = \sum_{i=1}^{k} N_k$. Let $R_i$ denote the sum of the ranks associated with the $i$th sample. It can be shown that the grand mean rank is $\frac{N+1}{2}$, whereas the sample mean rank for the $i$th sample is $\frac{R_i}{N_i}$. The test statistic takes the form of

$$\frac{12}{N(N+1)} \sum_{i=1}^{k} N_i \left( \frac{R_i}{N_i} - \frac{N+1}{2} \right)^2.$$

The term of $(\frac{R_i}{N_i} - \frac{N+1}{2})$ measures the deviation of the $i$th sample rank mean away from the grand rank mean. The term of $\frac{12}{N(N+1)}$ is the inverse of the variance of the total summation of ranks, and therefore it serves as a standardization factor. When $N$ is fairly large, the asymptotic distribution of Kruskal–Wallis statistic can be approximated by the chi-squared distribution with $k-1$ degrees of freedom.

## Tests for Factorial Design

### *Test for Main Effects*

In practice, data are often generated from experiments with several factors. To assess the individual factor effects in a nonparametric way, researchers could employ a rank transform method to test for the treatment effect of interest. Assume $X_{ijn} = \theta + \alpha_i + \beta_j + e_{ijn}$, while $i = 1, \ldots, I$ indexes for the blocks, $j = 1, \ldots, J$ indexes for the treatment levels, and $n = 1, \ldots, N$ indexes for the replicates. The null hypothesis to be tested is $H_0 : \beta_j = 0$, $j = 1, \ldots, J$, versus the alternative hypothesis $H_1$ : at least one $\beta_j \neq 0$. The noises $e_{ijn}$ are assumed to be independent and identically distributed with certain distribution $F$. The rank transform method proposed by W. J. Conover and Ronald L. Iman consists of replacing the observations by their ranks in the overall sample and then performing one of the standard ANOVA procedures on these ranks. Let $R_{ijn}$ be the rank corresponding to the observation $X_{ijn}$, and $\overline{R}_{ij\cdot} = 1/N \sum_n R_{ijn}, \overline{R}_{\cdot\cdot} = 1/(NI) \sum_i \sum_n R_{ijn}$. The Hora–Conover statistic proposed by Stephen C. Hora and Conover takes the form of

$$F = \frac{NI \sum_j (\overline{R}_{\cdot j\cdot} - \overline{R}_{\cdot\cdot\cdot})^2 / (J-1)}{\sum_i \sum_j \sum_k (R_{ijk} - \overline{R}_{ij\cdot})^2 / IJ(N-1)}.$$

When sample size is large, either the number of replicates per cell $N \to \infty$, or the number of blocks $I \to \infty$, the $F$ statistic has a limiting $\chi^2_{J-1}$ distribution. The $F$ statistic resembles the analysis of variance statistic in which the actual observation $X_{ijk}$s are replaced by $R_{ijk}$s. Such an analysis is easy to perform, as most software implement ANOVA procedures. This method has wide applicability in the analysis of experimental data because of its robustness and simplicity in use.

However, the Hora–Conover statistic $F$ cannot handle unbalanced designs that often arise in practice. Consider the following unbalanced design:

$$X_{ijn} = \theta + \alpha_i + \beta_j + \varepsilon_{ijn},$$

where $i = 1, \ldots, I$ and $j = 1, \ldots, J$ index levels for factors A and B, respectively, and $n = 1, \ldots, n_{ij}, N = \sum_{ij} n_{ij}$. We wish to test the hypothesis: $H_0 : \beta_j = 0 \forall j$ versus $H_1 : \beta_j \neq 0$ for some $j$. To address the problem of unbalance in designs, let us examine the composition of a traditional rank.

Define the function $u(x) = 1$ if $x \geq 0$; and $u(x) = 0$ if $x < 0$ and note that

$$R_{ijn} = \sum_{i'} \sum_{j'} \sum_{n'}^{n_{ij}} u(X_{ijn} - X_{i'j'n'}).$$

Thus, the overall rankings do not adjust for different sample sizes in unbalanced designs. To address this problem, we define the notion of a weighted rank.

**Definition**

Let $\Omega = \{X_{ijn}, i = 1, \ldots, I, j = 1, \ldots, J, n = 1, \ldots, n_{ij}\}$ be a collection of random variables. The weighted rank of $X_{ijn}$ within this set is

$$R^*_{ijn} = \frac{N}{IJ} \sum_{i'j'} \frac{1}{n_{i'j'}} \left[ \sum_{n'} u(X_{ijn} - X_{i'j'n'}) \right],$$

where $N = \sum_{ij} n_{ij}$.

Define $\mathbf{S}^*_N = [S^*_N(j), j = 1, \ldots, J]$ to be a vector of weighted linear rank statistics with components $S^*_N(j) = \frac{1}{IJ} \sum_i \frac{1}{n_{ij}} \sum_{n=1}^{n_{ij}} R^*_{ijn}$. Let $\overline{S}^*_N = \frac{1}{J} \sum_j S^*_N(j)$. Denote the covariance matrix of $\mathbf{S}^*_N$ as $\mathbf{\Sigma} = (\sigma_{b,b'})$, with $b, b' = 1, \ldots, J$. To estimate $\mathbf{\Sigma}$, We construct a variable $C^b_{ijn}$

$$\begin{aligned} C^b_{ijn} &= -\frac{1}{IJ^2 \rho_{ij}} (R^*_{ijn}/N), j \neq b \\ &\frac{J-1}{IJ^2 \rho_{ib}} (R^*_{ibn}/N), j = b \cdot \end{aligned} \tag{1}$$

Let $\hat{\sigma}_N(b, b') = \sum_i \sum_j \sum_n - (C^b_{ijm} - \bar{C}^b_{ij\cdot})^2$, and $\hat{\sigma}_N(b, b') = \sum_i \sum_j \sum_{n'} (C^b_{ijn} - \bar{C}^b_{ij\cdot})(C^{b'}_{ijn} - \bar{C}^{b'}_{ij\cdot})$. Let $\hat{\mathbf{\Sigma}}_N$ be the $J \times J$ matrix of $[\hat{\sigma}_N(b, b')], b, b' = 1, \ldots, J$. The fact that $R^*_{ijn}/N$ converges to $H(X_{ijn})$ almost surely leads to the fact that under $H_0, \forall j$, $\frac{1}{N}(\mathbf{\Sigma}_N - \hat{\mathbf{\Sigma}}_N) \to 0$ a.s. elementwise.

Construct a contrast matrix $\mathbf{A} = \mathbf{I}_J - \frac{1}{J}\mathbf{J}_J$. The generalized Hora–Conover statistic proposed by Xin Gao and Mayer Alvo for the main effects in unbalanced designs takes the form

$$T_M = (\mathbf{A}\mathbf{S}^*_N)' (\mathbf{A}\hat{\mathbf{\Sigma}}_N \mathbf{A}')^- (\mathbf{A}\mathbf{S}^*_N),$$

in which the general inverse of the covariance matrix is employed. The statistic $T_M$ is invariant with respect to choices of the general inverses. When the design is balanced, the test statistic is equivalent to the Hora–Conover statistic. This statistic $T_M$ converges to a central $\chi^2_{J-1}$ as $N \to \infty$.

### *Test for Nested Effects*

Often, practitioners might speculate that the different factors might not act separately on the response and interactions might exist between the factors. In light of such a consideration, one could study an unbalanced two-way layout with an interaction effect. Let $X_{ijn} (i = 1, \ldots, I;\ j = 1, \ldots, J, n = 1, \ldots, n_{ij})$ be a set of $N = \sum_i \sum_j n_{ij}$ independent random variables with the model

$$X_{ijn} = \theta + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijn}, \tag{2}$$

where $i$ and $j$ index levels for factors A and B, respectively. Assume $\sum_i \alpha_i = \sum_j \beta_j = \sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0$ and $\varepsilon_{ijn}$ are independent and identically distributed (i.i.d.) random variables with absolute continuous cumulative distribution function (cdf) $F$. Let $\delta_{ij} = \beta_j + \gamma_{ij}$. To test for nested effect, we consider testing $H_0 : \delta_{ij} = 0, \ \forall i$ and $j$, versus $H_1 : \delta_{ij} \neq 0$, for some $i$ and $j$. The nested effect can be viewed as the combined overall effect of the treatment either through its own main effect or through its interaction with the block factor.

The same technique of using weighted rank can be applied in testing for nested effects. Define $\mathbf{S}^*_N(i, j) = 1/n_{ij} \sum_n R^*_{ijn}$ and let $\mathbf{S}^*_N$ be the $IJ$ vector of $[S^*_N(i, j), 1 \leq i \leq I, 1 \leq j \leq J]$. Construct a contrast matrix $\mathbf{B} = \mathbf{I}_I \otimes (\mathbf{I}_J - \frac{1}{J}\mathbf{J}_J)$, such that the $ij$ element of $\mathbf{B}\mathbf{S}^*_N$ is $\mathbf{S}^*_N(i, j) - \frac{1}{J} \sum_{b=1}^{J} (i, b)$. Let $\mathbf{\Gamma}$ denote the covariance matrix of $\mathbf{S}^*_N$. To facilitate the estimation of $\mathbf{\Gamma}$, we define the following variables:

$$C^{(i,j)}(X_{abn}) = \begin{cases} \dfrac{N}{IJn_{ab}n_{ij}} \displaystyle\sum_{k=1}^{n_{ij}} u(X_{abn} X_{ijk}) \text{ for } (a, b) \neq (i, j). \\ l\dfrac{N}{IJn_{ij}} \displaystyle\sum_{(i',j')(i,j)} \frac{1}{n_{i'j'}} \sum_{n'=1}^{n_{i'j'}} u(X_{ijn} - X_{i'j'n'}) \text{ for } (a, b) = (i, j) \end{cases} \tag{3}$$

Let $\hat{\mathbf{\Gamma}}_N$ be the $IJ$ matrix with elements

$$\hat{\gamma}_N[(i,j),(i',j')] = \sum_{a,b,n}(C^{(i,j)})(X_{abn}) - \bar{C}^{(i,j)}[X_{ab\cdot}][C^{(i',j')}(X_{abn}) - \bar{C}^{(i',j')}(X_{ab})],$$

where $\bar{C}^{(i,j)}(X_{ab\cdot}) = \frac{1}{n_{ab}}\sum_n C^{(i,j)}(X_{abn})$. It can be proved that $1/\sqrt{N}(\hat{\gamma}_N - \gamma_N) \to 0$ almost surely elementwise. The proposed test statistic $T_N$ for nested effects takes the form

$$(\mathbf{BS}_N^*)'(\mathbf{B}\hat{\mathbf{\Gamma}}_N\mathbf{B}')^-(\mathbf{BS}_N^*).$$

Under $H_0 : \delta_{ij} = 0, \forall i, j$, the proposed statistic $T_N$ converges to a central $\chi^2_{I(J-1)}$ as $N \to \infty$.

## Tests for Pure Nonparametric Models

The previous discussion has been focused on the linear model with the error distribution unspecified. To reduce the model assumption, Michael G. Akritas, Steven F. Arnold, and Edgar Brunner have proposed a nonparametric framework in which the structures of the designs are no longer restricted to linear location models. The nonparametric hypotheses are formulated in terms of linear contrasts of normalized distribution functions. One advantage of the nonparametric hypotheses is that the parametric hypotheses in linear models are implied by the nonparametric hypotheses. Furthermore, the nonparametric hypotheses are not restricted to continuous distribution functions and therefore any models with discrete observations might also be included in this setup.

Under this nonparametric setup, the response variables in a two-way unbalanced layout with $I$ treatments and $J$ blocks can be described by the following model:

$$X_{ijn} : F_{ij}(x), i = 1,\ldots,I, j = 1,\ldots,J, \quad n = 1,\ldots,n_{ij}, \tag{2}$$

where $F_{ij}(x) = \frac{1}{2}[F^+_{ij}(x) + F^-_{ij}(x)]$ denotes the normalized-version of the distribution function, $F^+_{ij}(x) = P(X_{ijn} \le x)$ denotes the right continuous version, and $F^-_{ijn}(x) = P(X_{ijn} < x)$ denotes the left continuous version. The normalized version of the distribution function accommodates both ties and ordinal data. Compared with the classic ANOVA models, this nonparametric framework is different in two aspects: First, the normality assumption is relaxed; second, it not only includes the commonly used location models but also encompasses other arbitrary models with different cells having different distributions. Under this nonparametric setting, the hypotheses can be formulated in terms of linear contrasts of the distribution functions. According to Akritas and Arnold's method, $F_{ij}$ can be decomposed as follows:

$$F_{ij}(y) = M(y) + A_i(y) + B_j(y) + C_{ij}(y),$$

where $\sum_i A_i = \sum_j B_j = 0$, and $\sum_i C_{ij} = 0$, for all $j$, and $\sum_j C_{ij} = 0$, for all $i$. It follows that $M = \overline{F}_{\cdot\cdot}, A_i = \overline{F}_{i\cdot} - M, B_j = \overline{F}_{\cdot j} - M$, and $C_{ij} = F_{ij} - \overline{F}_{i\cdot} - \overline{F}_{\cdot j} + M$, where the subscript "$\cdot$" denotes summing over all values of the index. Denote the treatment factor as factor A and the block factor as factor B. The overall nonparametric hypotheses of no treatment main effects and no treatment simple factor effects are specified as follows:

$$H_0(A) : \overline{F}_{i\cdot} - \overline{F}_{\cdot\cdot} = 0, \forall i = 1,\ldots,I;$$
$$H_0(A|B) : F_{ij} - \overline{F}_{\cdot j} = 0, \forall i = 1,\ldots,I., \forall j = 1,\ldots,J.$$

The hypothesis $H_0(A|B)$ implies that the treatment has no effect on the response either through the main effects or through the interaction effects.

This framework especially accommodates the analysis of interaction effects in a unified manner. In literature, testing interactions using ranking methods have been a controversial issue for a long time. The problem pertaining to the analysis of interaction effects is because the interaction effects based on cell means can be artificially removed or introduced after certain nonlinear transformations. As rank statistics are invariant to nonlinear monotone transformations, they cannot be used to test for hypotheses that are not invariant to monotone transformations. To address this problem, Akritas and Arnold proposed to define nonparametric interaction effects in terms of linear contrasts of the distribution functions. Such nonparametric formulation of interaction effects is invariant to monotone transformations. The nonparametric hypothesis of

no interaction assumes the additivity of the distribution functions and is defined as follows:

$$H_0(AB): F_{ij} - \overline{F}_{i\cdot} - \overline{F}_{j\cdot} \cdot j + \overline{F}_{\cdot\cdot} = 0, \forall i = 1, \ldots, I., \quad \forall j = 1, \ldots, J.$$

This implies the distribution in the $(i,j)$th cell, $F_{ij}$, is a mixture of two distributions, one depending on $i$ and the other depending on $j$, and the mixture parameter is the same for all $(i,j)$. It is noted that all the nonparametric hypotheses bear analogous representation to their parametric counterparts except the parametric means are replaced by the distribution functions. Furthermore, all the nonparametric hypotheses are stronger than the parametric counterparts.

As no parameters are involved in the general model in Equation 4, the distribution functions $F_{ij}(x)$ can be used to quantify the treatment main effects, the treatment simple factor effects, and the interaction effects. To achieve this goal, Brunner and M. L. Puri considered the unweighted relative effects which take the form

$$\pi_{ij} = \int H^* dF_{ij}, i = 1, \ldots, I, j = 1, \ldots, J, \qquad (5)$$

where $H^*(x) = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} F_{ij}(x)$ is the average distribution function in the experiment. Denote $\mathbf{F} = (F_{11}, \ldots, F_{IJ})'$ as the vector of the distribution functions. In general, the measure $\pi_{ij}$ should be interpreted as the probability that the random variable generated from $F_{ij}$ will tend to be larger than a random variable generated from the average distribution function $H^*$. In the special case of shift models or by assuming noncrossing cumulative distribution functions, the relative effects $\pi_{ij}$ define a stochastic order by $F_{ij} <, =, < H^*$, according as $\pi_{ij} <, =,$ or $< \frac{1}{2}$. In practice, the definition of a relative effect is particularly convenient for ordinal data. As the differences of means can only be defined with metric scales, the parametric approach of comparing two treatments based on means is not applicable on ordinal scales. However, the points of an ordinal scale can be ordered by size, which can be used to form the estimate of the nonparametric relative effects.

Let $\boldsymbol{\pi} = (\pi_{11}, \ldots, \pi_{IJ})' = \int H^* d\mathbf{F}$ denote the vector of the relative effects. The relative effects $\pi_{ij}$ can be estimated by replacing the distribution functions $F_{ij}(x)$ by their empirical counterparts

$$\hat{F}_{ij}(x) = \frac{1}{2}[\hat{F}_{ij^+}(x) + \hat{F}_{ij^-}(x)] = \frac{1}{n_{ij}}\sum_{k=1}^{n_{ij}}(x - X_{ijk}),$$

where $c(u) = \frac{1}{2}[c^+(u) + c^-(u)]$ denotes the indicator function with $c^+(u) = 0$ or 1 depending on whether $u < 0$ or $\geq 0$, and $c^-(u) = 0$ or 1 depending on whether $u \leq 0$ or $> 0$. The vector of the empirical distribution functions is denoted by $\hat{\mathbf{F}} = (\hat{F}_{11}, \ldots, \hat{F}_{IJ})'$. The average empirical distribution function is denoted as $\hat{H}^*(x) = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{j}\hat{F}_{ij(x)}$. Then, the vector of the unweighted relative effects is estimated unbiasedly by $\hat{\boldsymbol{\pi}} = \int H^* d\hat{\mathbf{F}} = \frac{1}{N}\bar{\mathbf{R}}^*$, where $N = \sum_i\sum_j n_{ij}$, $\bar{\mathbf{R}}^* = (\bar{R}^*_{11\cdot}, \ldots, \bar{R}^*_{IJ\cdot})', \bar{R}^*_{ij\cdot} = \frac{1}{n_{ij}}\sum_{k=1}^{n_{ij}} R^*_{ijk}$, and $R^*_{ijk} = \frac{N}{IJ}\sum_{i'=1}^{I}\frac{1}{n^{i'j'}}\sum_{k'=1}^{n'_{i'j'}} - X_{i'j'k'}$. The latter is the weighted rank of the observation $X_{ijk}$ among all the observations proposed by Sebastian Domhof, Gao, and Alvo.

Focusing on the problem of hypothesis testing in unbalanced two-way layouts, with $H_0: \mathbf{CF} = \mathbf{0}$, versus $H_0: \mathbf{CF} \neq \mathbf{0}$. and $\mathbf{C}$ being a $IJ$ by $IJ$ contrast matrix. It can be shown that the testing main effects, nested effects, or interaction effects can all be expressed in such a general form. Assume that as $\min\{n_{ij}\} \to \infty$, $\lim_{N\to\infty} n_{ij}/N = \lambda_{ij}$, and $\sigma^2_{ij} = \mathrm{var}[H^*(X_{ijk})] > 0$, for all $i$ and $j$. Then, an application of the Lindeberg-Feller theorem yields $\sqrt{N}(\mathbf{C}\hat{\boldsymbol{\pi}}) \xrightarrow{d} N(\mathbf{0}, \mathbf{CVC}')$, with $\mathbf{V} = \mathrm{diag}, (\frac{1}{\lambda_{11}}\sigma^2_{11}, \ldots, \frac{1}{\lambda_{IJ}}\sigma^2_{IJ})$. The asymptotic variance $\sigma^2_{ij}$, $i = 1, \ldots, I$, and $j = 1, \ldots, J$, can be estimated from the pseudo ranks as $\hat{\sigma}^2_{ij} = \frac{1}{N^2(n_{ij}-1)}\sum_{k=1}^{n_{ij}}(R^*_{ijk} - \overline{R}^*_{ij\cdot})^2$, and the estimated covariance matrix takes the form of $\hat{\mathbf{V}} = \mathrm{diag}(\frac{1}{\lambda_{11}}\hat{\sigma}^2_{11}, \ldots, \frac{1}{\lambda_{IJ}}\hat{\sigma}^2_{IJ})$. The fact that $R^*_{ijk}/N$ converges to $H^*(X_{ijk})$ almost surely leads to the result that $\mathbf{V} - \hat{\mathbf{V}} \xrightarrow{a.s.} \mathbf{0}$ elementwise. Therefore, we consider the test statistic $T = \sqrt{N}\hat{\boldsymbol{\pi}}'\mathbf{C}'(\mathbf{C}\hat{\mathbf{V}}\mathbf{C}')^-\mathbf{C}\hat{\boldsymbol{\pi}}$, where $(\mathbf{C}\hat{\mathbf{V}}\mathbf{C}')^-$ denotes the general inverse of $(\mathbf{C}\hat{\mathbf{V}}\mathbf{C}')$. According to Slutsky's theorem, because $\hat{\mathbf{V}}$ is consistent, we have $T \xrightarrow{d} \chi^2_f$, where the degrees of freedom $f = \mathrm{rank}(C)$.

## Conclusions

Nonparametric methods have wide applicabilities in the analysis of experimental data because of their robustness to the underlying distributions and insensitivity to potential outliers. In this entry, various methods to test for different effects in a factorial design have been reviewed. It is worthy to emphasize that in the situation where the normality assumption is violated, nonparametric methods are powerful alternative approaches to make sounded statistical inference for experimental data.

*Xin Gao*

***See also*** Distribution; Normal Distribution; *Nonparametric Statistics for the Behavioral Sciences;* Null Hypothesis; Research Hypothesis

### Further Readings


Akritas, M. G., & Arnold, S. F. (1994). Fully nonparametric hypotheses for factorial designs I: Multivariate repeated measures designs. *Journal of American Statistical Association,* 89, 336–343.

Akritas, M., Arnold, S., & Brunner, E. (1997). Nonparametric hypotheses and rank statistics for unbalanced factorial designs. *Journal of American Statistical Association,* 92, 258–265.

Akritas, M. G., & Brunner, E. (1997). A unified approach to rank tests in mixed models. *Journal of Statistical Planning and Inferences,* 61, 249–277.

Brunner, E., Puri, M. L., & Sun, S. (1995). Nonparametric methods for stratified two-sample designs with application to multi-clinic trials. *Journal of American Statistical Association,* 90, 1004–1014.

Brunner, E., & Munzel, U. (2002). *Nichtparametrische datenanalyse*. Heidelberg, Germany: Springer-Verlag.

Conover, W. J., & Iman, R. L. (1976). On some alternative procedures using ranks for the analysis of experimental designs. *Communication in Statistics, A5,* 1349–1368.

Domhof, S. (2001). *Nichtparametrische relative effekte* (Unpublished dissertation). University of Göttingen, Göttingen, Germany.

Gao, X., & Alvo, M., (2005). A unified nonparametric approach for unbalanced factorial designs. *Journal of American Statistical Association,* 100, 926–941.

Hora, S. C., & Conover, W. J. (1984). The F statistic in the two-way layout with rank-score transformed data. *Journal of the American Statistical Association,* 79, 668–673.

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of American Statistical Association,* 47, 583–621.

Mann, H., & Whitney, D. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics,* 18, 50–60.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin,* 1, 80–83.


# *Nonparametric Statistics for the Behavioral Sciences*

Sidney Siegel (January 4, 1916–November 29, 1961) was a psychologist trained at Stanford University. He spent nearly his entire career as a professor at Pennsylvania State University. He is known for his contribution to nonparametric statistics, including the development with John Tukey of the Siegel–Tukey test—a test for differences in scale between groups. Arguably, he is most well known for his book, *Nonparametric Statistics for the Behavioral Sciences,* the first edition of which was published by McGraw-Hill in 1956. After Siegel's death, a second edition was published (1988) adding N. John Castellan, Jr., as coauthor. *Nonparametric Statistics for the Behavior Sciences* is the first text to provide a practitioner's introduction to nonparametric statistics. By its copious use of examples and its straightforward "how to" approach to the most frequently used nonparametric tests, this text was the first accessible introduction to nonparametric statistics for the nonmathematician. In that sense, it represents an important step forward in the analysis and presentation of non-normal data, particularly in the field of psychology.

The organization of the book is designed to assist the researcher in choosing the correct nonparametric test. After the introduction, the second chapter introduces the basic principles of hypothesis testing, including the definitions of: the null and alternative hypothesis, the size of the test, Type I and Type II errors, power, sampling distributions, and the decision rule. Chapter 3 describes the factors that influence the choice of correct test. After explaining some common parametric assumptions and the circumstances under which nonparametric

tests should be used, the text gives a basic outline of how the proper statistical test should be chosen. Tests are distinguished from one another in two important ways: First, tests are distinguished by their capability of analyzing data of varying levels of measurement. For example, the $\chi 2$ goodness-of-fit test can be applied to nominal data, whereas the Kolmogorov–Smirnov requires at least the ordinal level of measurement. Second, tests are distinguished in terms of the type of samples to be analyzed. For example, two-sample paired tests are distinguished from tests applicable to $k$ independent samples, which are distinguished tests of correlation, and so on. Tests included in the text include the following: the binomial test, the sign test, the signed-rank test, tests for data displayed in two-way tables, the Mann–Whitney $U$ test, the Kruskal–Wallis test, and others. Also included are extensive tables of critical values for the various tests discussed in the text.

Because nonparametric tests make fewer assumptions than parametric tests, they are generally less powerful than the parametric alternatives. The text compares the various tests presented with their parametric analogues in terms of power efficiency. Power efficiency is defined to be the percent decrease in sample size required for the parametric test to achieve the same power as that of the nonparametric test when the test is performed on data that do, in fact, satisfy the assumptions of the parametric test.

This work is important because it seeks to present nonparametric statistics in a way that is "completely intelligible to the reader whose mathematical training is limited to elementary algebra" (Siegel, 1956, p. 4). It is replete with examples to demonstrate the application of these tests in contexts that are familiar to psychologists and other social scientists. The text is organized so that the user, knowing the specific level of measurement and type(s) of samples being analyzed, can immediately identify several nonparametric tests that might be applied to his or her data.

Included in each test is a description of its function (under what circumstances this particular test should be used), rationale, and method (a heuristic description of why the test works and how the test statistic is calculated) including any modifications that exist and the procedure for dealing with ties, both large and small sample examples, a numbered list of steps for performing the test, and other references for a more in-depth description of the test.

*Gregory Michaelson and Michael Hardin*

***See also*** Distribution; Nonparametric Statistics; Normal Distribution; Null Hypothesis; Research Hypothesis

**Further Readings**

Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.

# Nonprobability Sampling

The two kinds of sampling techniques are probability and nonprobability sampling. Probability sampling is based on the notion that the people or events chosen are selected because they are representative of the entire population. Nonprobability refers to procedures in which researchers select their sample elements not based on a predetermined probability. This entry examines the application, limitations, and utility of nonprobability sampling procedures. Conceptual and empirical strategies to use nonprobability sampling techniques more effectively are also discussed.

## Sampling Procedures

### *Probability Sampling*

There are many different types of probability sampling procedures. More common ones include simple, systematic, stratified, multistage, and cluster sampling. Probability sampling allows one to have confidence that the results are accurate and unbiased, and it allows one to estimate how precise the data are likely to be. The data from a properly drawn sample are superior to data drawn from individuals who just show up at a meeting or perhaps speak the loudest and convey their personal thoughts and sentiments. The critical issues in sampling include whether to use a probability sample, the sampling frame (the set of people that have a chance of being selected and how well it corresponds to the population studied), the size of the sample, the sample design (particularly the

strategy used to sample people, schools, households, etc.), and the response rate. The details of the sample design, including size and selection procedures, influence the precision of sample estimates regarding how likely the sample is to approximate population characteristics. The use of standardized measurement tools and procedures also helps to assure comparable responses.

### *Nonprobability Sampling*

Nonprobability sampling is conducted without the knowledge about whether those chosen in the sample are representative of the entire population. In some instances, the researcher does not have sufficient information about the population to undertake probability sampling. The researcher might not even know who or how many people or events make up the population. In other instances, nonprobability sampling is based on a specific research purpose, the availability of subjects, or a variety of other nonstatistical criteria. Applied social and behavioral researchers often face challenges and dilemmas in using a random sample, because such samples in a real-world research are "hard to reach" or not readily available. Even if researchers have contact with hard to reach samples, they might be unable to obtain a complete sampling frame because of peculiarities of the study phenomenon. This is especially true when studying vulnerable or stigmatized populations, such as children exposed to domestic violence, emancipated foster care youth, or runaway teenagers. Consider for instance the challenges of surveying adults with the diagnosis of paranoid personality disorder. This is not a subgroup that is likely to agree to sign a researcher's informed consent form, let alone complete a lengthy battery of psychological instruments asking a series of personal questions.

Applied researchers often encounter other practical dilemmas when choosing a sampling method. For instance, there might be limited research resources. Because of limitations in funding, time, and other resources necessary for conducting large-scale research, researchers often find it difficult to use large samples. Researchers employed by a single site or agency might be unable to access subjects served by other agencies located in other sites. It is not a coincidence that recruiting study subjects from a single site or agency is one of the most popular methods among studies using nonprobability procedures. The barriers preventing a large-scale multisite collaboration among researchers can be formidable and difficult to overcome.

## Statistical Theories About Sampling Procedures

Because a significant number of studies employ nonprobability samples and at the same time apply inferential statistics, it is important to understand the consequences. In scientific research, there are many reasons to observe elements of a sample rather than a population. Advantages of using sample data include reduced cost, greater speed, greater scope, and greater accuracy. However, there is no reason to use a biased sample that does not represent a target population. Scientists have given this topic a rigorous treatment and developed several statistical theories about sampling procedures. Central limit theorem, which is usually given in introductory statistics courses, forms the foundation of probability-sampling techniques. At the core of this theorem are the proven relationships between the mean of a sampling distribution and the mean of a population, between the standard deviation of a sampling distribution (known as standard error) and the standard deviation of a population, and between the normal sampling distribution and the possible non-normal population distribution.

Statisticians have developed various formulas for estimating how closely the sample statistics are clustered around the population true values under various types of sampling designs, including simple random sampling, systematic sampling, stratified sampling, clustered sampling, and multistage clustered and stratified sampling. These formulas become the yardstick for determining adequate sample size. Calculating the adequacy of probabilistic samples size is generally straightforward and can be estimated mathematically based on preselected parameters and objectives (i.e., $x$ statistical power with $y$ confidence intervals). In practice, however, sampling error—the key component required by the formulas for figuring out a needed sample size—is often unknown to researchers. In such instances, which often involve quasi-experimental designs, Jacob Cohen's framework of

statistical power analysis is employed instead. This framework concerns the balance among four elements of a study: sample size, effect size or difference between comparison groups, probability of making a Type I error, and probability of denying a false hypothesis or power. Studies using small nonprobability samples, for example, are likely to have an inadequate power (significantly below .85 convention indicating an adequate power). As a consequence, studies employing sophisticated analytical models might not meet the required statistical criteria. The ordinary least-square regression model, for example, makes five statistical assumptions about data, and most of the assumptions require a randomized process for data gathering. Violating statistical assumptions in a regression analysis refers to the presence of one or more detrimental problems such as heteroscedasticity, autocorrelation, non-normality, multicollinearity, and others. Multicollinearity problems are particularly likely to occur in nonprobability studies in which data were gathered through a sampling procedure with hidden selection bias and/or with small sample sizes. Violating statistical assumptions might increase the risk of producing biased and inefficient estimates of regression coefficients and exaggerated $R^2$.

## Guidelines and Recommendations

This brief review demonstrates the importance of using probability sampling; however, probability sampling cannot be used in all instances. Therefore, the following questions must be addressed: Given the sampling dilemmas, what should researchers do? How can researchers using nonprobability sampling exercise caution in reporting findings or undertake remedial measures? Does nonprobability sampling necessarily produce adverse consequences? It is difficult to offer precise remedial measures to correct the most commonly encountered problems associated with the use of nonprobability samples because such measures vary by the nature of research questions and type of data researchers employ in their studies. Instead of offering specific measures, the following strategies are offered to address the conceptual and empirical dilemmas in using nonprobability samples.

George Judge et al. caution researchers to be aware of assumptions embedded in the statistical models they employ, to be sensitive to departures of data from the assumptions, and to be willing to take remedial measures. John Neter et al. recommend that researchers always perform diagnostic tests to investigate departures of data from the statistical assumptions and take corrective measures if detrimental problems are present. In theory, all research should use probabilistic sampling methodology, but in practice this is difficult especially for hard to reach, hidden, or stigmatized populations. Much of social science research can hardly be performed in a laboratory. It is important to stress that the results of the study are meaningful if they are interpreted appropriately and used in conjunction with statistical theories. Theory, design, analysis, and interpretation are all connected closely.

Researchers are also advised to study compelling populations and compelling questions. This most often involves purposive samples in which the research population has some special significance. Most commonly used samples, particularly in applied research, are purposive. Purposive sampling is more applicable in exploratory studies and studies contributing new knowledge. Therefore, it is imperative for researchers to conduct a thorough literature review to understand the "edge of the field" and whether the study population or question is a new or significant contribution. How does this study contribute uniquely to the existing research knowledge? Purposive samples are selected based on a predetermined criteria related to the research. Research that is field oriented and not concerned with statistical generalizability often uses nonprobabilistic samples. This is especially true in qualitative research studies. Adequate sample size typically relies on the notion of "saturation," or the point in which no new information or themes are obtained from the data. In qualitative research practice, this can be a challenging determination.

Researchers should also address subject recruitment issues to reduce selection bias. If possible, researchers should use consecutive admissions including all cases during a representative time frame. They should describe the population in greater detail to allow for cross-study comparisons. Other researchers will benefit from additional data

and descriptors that provide a more comprehensive picture of the characteristics of the study population. It is critical in reporting results (for both probability and nonprobability sampling) to tell the reader who was or was not given a chance to be selected. Then, to the extent that is known, researchers should tell the reader how those omitted from the study the same or different from those included. Conducting diagnostics comparing omitted or lost cases with the known study subjects can help in this regard. Ultimately, it is important for researchers to indicate clearly and discuss the limits of generalizability and external validity.

In addition, researchers are advised to make efforts to assure that a study sample provides adequate statistical power for hypothesis testing. It has been shown that other things being equal, a large sample always produces more efficient and unbiased estimates about population true parameters than a small sample. When the use of a nonprobability sample is inevitable, researchers should carefully weigh the pros and cons that are associated with different study designs and choose a sample size that is as large as possible.

Another strategy is for researchers to engage in multiagency research collaborations that generate samples across agencies and/or across sites. In one study, because of limited resources, Brent Benda and Robert Flynn Corwyn found it unfeasible to draw a nationally representative sample to test the mediating versus moderating effects of religion on crime in their study. To deal with the challenge, they used a comparison between two carefully chosen sites: random samples selected from two public high schools involving 360 adolescents in the inner city of a large east coast metropolitan area and simple random samples involving 477 adolescents from three rural public high schools in an impoverished southern state. The resultant data undoubtedly had greater external validity than studies based on either site alone.

If possible, researchers should use national samples to run secondary data analyses. These databases were created by probability sampling and are deemed to have a high degree of representativeness and other desirable properties. The drawback is that these databases are likely to be useful for only a minority of research questions.

Finally, does nonprobability sampling necessarily produce adverse consequences? Shenyang Guo and David L. Hussey have shown that a homogeneous sample produced by nonprobability sampling is better than a less homogeneous sample produced by probability sampling in prediction. Remember, regression is a leading method used by applied researchers employing inferential statistics. Regression-type models also include simple linear regression, multiple regression, logistic regression, structural equation modeling, analysis of variance (ANOVA), multivariate analysis of variance (MANOVA), and analysis of covariance (ANCOVA). In a regression analysis, a residual is defined as the difference between the observed value and model-predicted value of the dependent variable. Researchers are concerned about this measure because it is the model with the smallest sample residual that gives the most accurate predictions about sample subjects. Statistics such as Theil's U can gauge the scope of sample residuals, which is a modified version of root-mean-square error measuring the magnitude of the overall sample residual. The statistic ranges from zero to one, with a value closer to zero indicating a smaller overall residual. In this regard, nonprobability samples can be more homogeneous than a random sample. Using regression coefficients (including an intercept) to represent study subjects, it is much easier to obtain an accurate estimate for a homogeneous sample than for a heterogeneous sample. This consequence, therefore, is that small homogeneous samples generated by a nonprobability sampling procedure might produce more accurate predictions about sample subjects. Therefore, if the task is not to infer statistics from sample to population, using a nonprobability sample is a better strategy than using a probability sample.

With the explosive growth of the World Wide Web and other new electronic technologies such as search monkeys, nonprobability sampling remains an easy way to obtain feedback and collect information. It is convenient, verifiable, and low cost, particularly when compared with face-to-face paper and pencil questionnaires. Along with the benefits of new technologies, however, the previous cautions apply and might be even more important given the ease with which larger samples might be obtained.

*David L. Hussey*

*See also* Naturalistic Inquiry; Probability Sampling; Sampling; Selection

### Further Readings

Benda, B. B., & Corwyn, R. F. (2001). Are the effects of religion on crime mediated, moderated, and misrepresented by inappropriate measures? *Journal of Social Service Research, 27,* 57–86.
Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: Wiley.
Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.
Guest, G., Bunce, A., & Johnson, L. (2006). How many interviews are enough? An experiment with data saturation and variability. *Field Methods, 18,* 59–82.
Guo, S., & Hussey, D. (2004). Nonprobability sampling in social work research. *Journal of Social Service Research, 30,* 1–18.
Judge, G. G., Griffiths, W. E., Hill, R. C., Lutkepohl, H., & Lee, T. C. (1985). *The theory and practice of econometrics* (2nd ed.). New York: Wiley.
Kennedy, P. (1985). *A guide to econometrics* (2nd ed.). Cambridge: MIT Press.
Kish, L. (1965). *Survey sampling*. New York: Wiley.
Mech, E. V., & Che-Man Fung, C. (1999). Placement restrictiveness and educational achievement among emancipated foster youth. *Research on Social Work Practice,* 9, 213–228.
Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear regression models* (3rd ed.). Chicago: Irwin.
Nugent, W. R., Bruley, C., & Allen, P. (1998). The effects of aggression replacement training on antisocial behavior in a runaway shelter. *Research on Social Work Practice,* 8, 637–656.
Peled, E., & Edleson, J. L. (1998). Predicting children's domestic violence service participation and completion. *Research on Social Work Practice,* 8, 698–712.
Rubin, A., & Babbie, E. (1997). *Research methods for social work* (3rd ed.). Pacific Grove, CA: Brooks/Cole.
Theil, H. (1966). *Applied economic forecasting*. Amsterdam, the Netherlands: Elsevier.

## Nonsignificance

This entry defines nonsignificance within the context of null hypothesis significance testing (NHST), the dominant scientific statistical method for making inferences about populations based on sample data. Emphasis is placed on the three routes to nonsignificance: a real lack of effect in the population; failure to detect a real effect because of an insufficiently large sample; or failure to detect a real effect because of a methodological flaw. Of greatest importance is the recognition that nonsignificance is not affirmative evidence of the absence of an effect in the population.

Nonsignificance is the determination in NHST that no statistically significant effect (e.g., correlation, difference between means, and dependence of proportions) can be inferred for a population. NHST typically involves statistical testing (e.g., *t* test) performed on a sample to infer whether two or more variables are related in a population. Studies often have a high probability of failing to reject a false null hypothesis (i.e., commit a Type II, or false negative, error), thereby returning a nonsignificant result even when an effect is present in the population.

In its most common form, a null hypothesis posits that the means on some measurable variable for two groups are equal to each other. A statistically significant difference would indicate that the probability that a true null hypothesis is erroneously rejected (Type I, or false positive, error) is below some desired threshold ($\alpha$), which is typically .05. As a result, statistical significance refers to a conclusion that there likely is a difference in the means of the two population groups. In contrast, nonsignificance refers to the finding that the two means do not significantly differ from each other (a failure to reject the null hypothesis). Importantly, nonsignificance does not indicate that the null hypothesis is true, it only indicates that one cannot rule out chance and random variation to explain observed differences. In this sense, NHST is analogous to an American criminal trial, in which there is a presumption of innocence (equality), the burden of proof is on demonstrating guilt (difference), and a failure to convict (reject the null hypothesis) results only in a verdict of "not guilty" (not significant), which does not confer innocence (equality).

Nonsignificant findings might reflect accurately the absence of an effect or might be caused by a research design flaw leading to low statistical power and a Type II error. Statistical power is defined as the probability of detecting an existing effect (rejecting a false null hypothesis) and might

be calculated ex ante given the population effect size (or an estimate thereof), the desired significance level (e.g., .05), and the sample size.

Type II errors resulting from insufficient statistical power can result from several factors. First, small samples yield lower power because they are simply less likely than large samples to be representative of the population, and they lead to larger estimates of the standard error. The standard error is estimated as the sample standard deviation divided by the square root of the sample size. Therefore, the smaller the sample, the bigger the estimate of the standard error will be. Because the standard error is the denominator in significance test equations, the bigger it is, the less likely the test statistic will be large enough to reject the null hypothesis. Small samples also contribute to nonsignificance because sample size (specifically, the degrees of freedom that are derived from it) is an explicit factor in calculations of significance levels ($p$ values). Low power can also result from imprecise measurement, which might result in excessive variance. This too will cause the denominator in the test statistic calculation to be large, thereby underestimating the magnitude of the effect.

Type II error can also result from flawed methodology, wherein variables are operationalized inappropriately. If variables are not manipulated or measured well, the real relationship between the intended variables will be more difficult to discern from the data. This issue is often referred to as *construct validity.* A nonrepresentative sample, even if it is large, or a misspecified model might also prevent the detection of an existing effect.

To reduce the likelihood of nonsignificance resulting from Type II errors, an a priori power analysis can determine the necessary sample size to provide the desired likelihood of rejecting the null hypothesis if it is false. The suggested convention for statistical power is .8. Such a level would allow a researcher to say with 80% confidence that no Type II error had been committed and, in the event of nonsignificant findings, that no effect exists.

Some have critiqued the practice of reporting significance tests alone, given that with a .05 criterion, determining that a result of .049 is statistically significant, whereas one of .051 is not artificially dichotomizes the determination of significance. An overreliance on statistical significance also results in a bias among published research for studies with significant findings, contributing to a documented upward bias in effect sizes in published studies. Directional hypotheses might also be an issue. For accurate determination of significance, directionality should be specified in the design phase, as directional hypotheses (e.g., predicting that one particular mean will be higher than the other) have twice the statistical power of nondirectional hypotheses, provided the results are in the hypothesized direction.

Many of these problems can be avoided with a careful research design that incorporates a sufficiently large sample based on an a priori power analysis. Other suggestions include reporting effect sizes (e.g., Cohen's $d$) and confidence intervals to convey more information about the magnitude of effects relative to variance and how close the results are to being determined significant. Additional options include reporting the power of the tests performed or the sample size needed to determine significance for a given effect size. Additionally, meta-analysis—combining effects across multiple studies, even those that are nonsignificant—can provide more powerful and reliable assessments of relations among variables.

*Christopher Finn and Jack Glaser*

***See also*** Null Hypothesis; Power; Power Analysis; Significance, Statistical; Type II Error

**Further Readings**

Berry, E. M., Coustere-Yakir, C., & Grover, N. B. (1998). The significance of non-significance. *QJM, 91,* 647–653.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Cohen, J. (1994). The earth is round. *American Psychologist, 49,* 997–1003.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin,* 86, 638–641.

## Normal Distribution

The normal distribution, which is also called a Gaussian distribution, bell curve, or normal curve, is commonly known for its bell shape (see

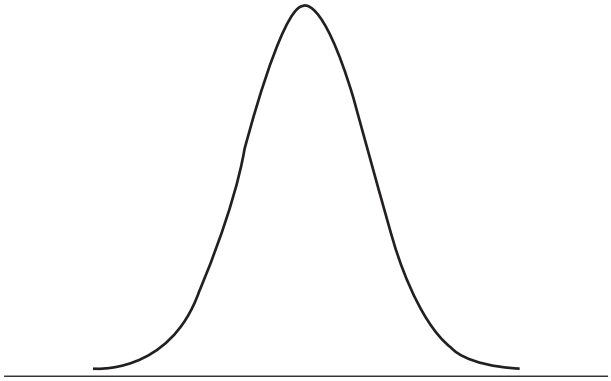


**Figure 1** The Normal Distribution

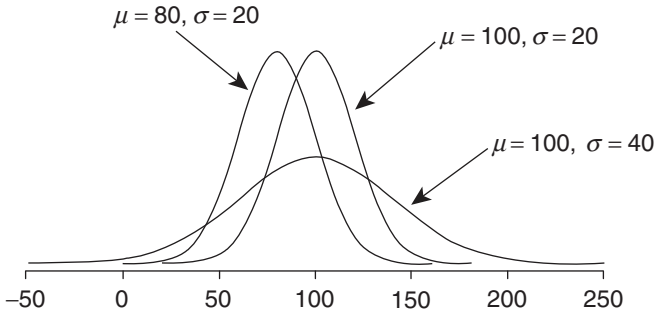



**Figure 2** Examples of Three Different Normal Distributions

Figure 1) and is defined by a mathematical formula. It is a member of families of distributions such as exponential, monotone likelihood ratio, Pearson, stable, and symmetric power. Many biological, physical, and psychological measurements, as well as measurement errors, are thought to approximate normal distributions. It is one of the most broadly used distributions to describe continuous variables.

The normal curve has played an essential role in statistics. Consequently, research and theory have grown and evolved because of the properties of the normal curve. This entry first describes the characteristics of the normal distribution, followed by a discussion of its applications. Lastly, this entry gives a brief history of the normal distribution.

## Characteristics of the Normal Distribution

The normal distribution has several properties, including the following:

- The curve is completely determined by the mean (average) and the standard deviation (the spread about the mean, or girth).
- The mean is at the center of this symmetrical curve (which is also the peak or maximum ordinate of the curve).
- The distribution is unimodal; the mean, median, and mode are the same.
- The curve is asymptotic. Values trail off symmetrically from the mean in both directions, indefinitely (the tails never touch the $x$ axis) and form the two tails of the distribution.
- The curve is infinitely divisible.
- The skewness and kurtosis are both zero.

Different normal density curves exist because a normal distribution is determined by the mean and the standard deviation. Figure 2 presents three normal distributions, one with a mean of 100 and a standard deviation of 20, another normal distribution with a mean of 100 and a standard deviation of 40, and a normal distribution with a mean of 80 and a standard deviation of 20. The curves differ with respect to their spread and their height. There can be an unlimited number of normal distributions because there are an infinite number of means and standard deviations, as well as combinations of those means and standard deviations.

Although sometimes labeled as a "bell curve," the curve does not always resemble a bell shape (e.g., the distribution with a mean of 100 and standard deviation of 40 is flatter in this instance because of the scale being used for the $x$ and $y$ axes). Also, not all bell-shaped curves are normal distributions. What determines whether a curve is a normal distribution is not dependent on its appearance but on its mathematical function.

The normal distribution is a mathematical curve defined by the probability density function (PDF):

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2},$$

where

$f(x)$ = the density of the function or the height of the curve (usually plotted on the $y$ axis) for

a particular variable $x$ (usually plotted on the $x$ axis) with a normal distribution

$\sigma$ = standard deviation of the distribution

$\pi$ = the constant 3.1416

$e$ = base of Napierian logarithms, 2.7183

$\mu$ = mean of the distribution

This equation dictates the shape of the normal distribution. Normal distributions with the same mean and standard deviation would be identical in form. The curve is symmetrical about its mean because $(x - \mu)$ is squared. Furthermore, because the exponent is negative, the more $x$ deviates from the mean (large positive number or large negative number), $f(x)$ becomes very small, infinitely small but never zero. This explains why the tails of the distribution would never touch the $x$ axis.

The total area under the normal distribution curve equals one (100%). Mathematically, this is represented by the following formula:

$$\int_{-\infty}^{+\infty} f(x)dx = 1.$$

A probability distribution of a variable $X$ can also be described by its cumulative distribution function (CDF). The mathematical formula for the CDF is:

$$F(x) = p(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt.$$

$F(x)$ gives the probability that any randomly chosen variable $X$, with a normal distribution, is less than or equal to $x$, the value of interest. The normal CDF curve is presented in Figure 3 along with the normal PDF curve. The value of the CDF at a point on the $x$ axis equals the area of the PDF up to that same value. CDF values range from 0 to 1.

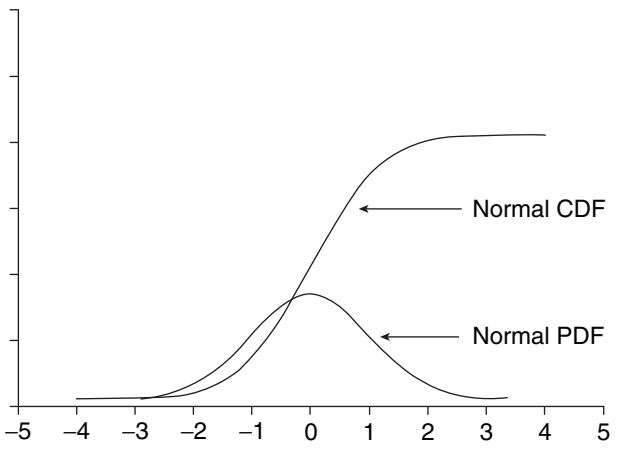



**Figure 3** The Normal CDF and the Normal PDF Curves

## Applications

The term *normal* was coined to reflect that this probability distribution is commonly found and not that other probability distributions are somehow not "normal." Although at one time it was extolled as the distribution of all traits, it is now recognized that many traits are best described by other distributions. Exponential distributions, Lévy distributions, and Poisson distributions are some examples of the numerous other distributions that exist.

Some variables that can be said to be approximated by the normal curve include height, weight, personality traits, intelligence, and memory. They are approximated because their real distributions would not be identical to that of a normal distribution. For instance, the real distributions of these variables might not be perfectly symmetrical, their curves might bend slightly, and their tails do not go to infinity. (The tails of the normal distribution go to infinity on the $x$ axis. This observation is not possible for the variables mentioned previously; e.g., certain heights for a human being would be impossible.) The normal distribution reflects the best shape of the data. It is an idealized version of the data (see Figure 4), is described by a mathematical function (the PDF), and because many variables are thought to approximate the normal distribution, it can be used as a model to describe the distribution of the data. Thus, it is possible to take advantage of many of the normal curve's strengths.

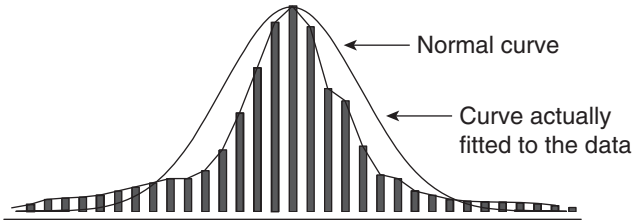



**Figure 4** An Example of a Normal Curve and a Curve Fitted to the Data

One strength of these normal curves is that they all share a characteristic: probabilities can be identified. However, because normal distributions can take on any mean and/or standard deviation, calculations for each normal distribution would be required to determine the area or probability for a given score or the area between two scores. To avoid these lengthy calculations, the values are converted to standard scores (also called standard units or $z$ scores). Many statistics textbooks present a table of probabilities that is based on a distribution called the standard normal distribution. The standard normal curve consists of a mean ($\mu$) of zero and a standard deviation ($\sigma$) of one. The probabilities for the standard normal distribution have been calculated and are known, and because the values of any normal distribution can be converted to a standard score, the probabilities for each normal curve do not need to be calculated but can be found through their conversion to $z$ scores. The density function of the standard normal distribution is

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

Figure 5 presents a standard normal curve; 68.27% of the values in a particular data set lie between $-1\,\sigma$ and $+1\,\sigma$, 34.13% lie between the mean and $+1\,\sigma$, 34.13% lie between the mean and $-1\,\sigma$), 95.45% of the values lie between $-2\,\sigma$ and $+2\,\sigma$, 13.59% of the cases lie between $+1\,\sigma$ and $+2\,\sigma$, and 13.59% lie between $-2\,\sigma$ and $-1\,\sigma(2 \times 13.59\%) + (2 \times 34.13\%) = 95.45\%$, and 99.97% of all values lie between $-3\,\sigma$ and $+3\,\sigma$. Because 99.97% of all values lie between $-3$ and $+3$ standard deviations, a value found outside of this range would be rare.

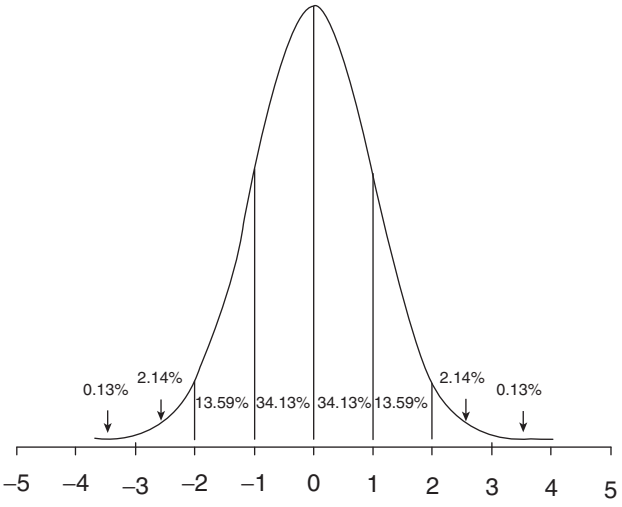



**Figure 5** The Standard Normal Distribution and Its Percentiles

Given that all normal curves share the same mathematical function and have the same area under the curve, variables that are normally distributed can be converted to a standard normal distribution using the following formula:

$$z = \frac{(X - \mu)}{\sigma},$$

where

$z$ = standard score on $x$ and represents how many standard deviations a raw score ($x$) is above or below the mean

$X$ = score of interest for the normally distributed data

$\mu$ = mean for the normally distributed data

$\sigma$ = standard deviation for the normally distributed data

In standardizing a set of scores they do not become normal. In other words, the shape of a distribution of scores does not change (i.e., does not become normal) by converting them to $z$ scores.

Once scores are standardized, the percentile (proportion of scores) for a given data point/score can be determined easily. To illustrate, if the average score for an intelligence measure is 100 and the standard deviation is 10, then using the above formula and the table of the area of the standard

normal distribution, it is possible to determine percentages (proportions) of cases that have scored less than 120. First, the calculation of the $z$ score is required:

$$z = \frac{(X - \mu)}{\sigma} = \frac{(120 - 100)}{10} = 2.$$

A $z$ score of +2 means that the score of 120 is 2 standard deviations above the mean. To determine the percentile for this $z$ score, this value has to be looked up in a table of standardized values of the normal distribution (i.e., a $z$ table). The value of 2 is looked up in a standard normal-curve areas table (not presented here, but can be found in most statistics textbooks) and the corresponding value of .4772 is found (this is the area between the mean and $z$ value of 2). This means that the probability of observing a score between 100 (the mean in this example) and 120 (the score of interest in this example) is 47.72. The standard normal distribution is symmetrical around its mean so 50% of all scores are 100 and less. To determine what proportion of individuals score below 120, the value below 100 has to be added to the value between 100 and 120. Therefore, 50% is added to 47.72% resulting in 97.72%. Thus, 97.72% of individuals are expected to score worse than or equal to 120. Conversely, if interested in determining the proportion of individuals who would score better than 120, .4772 would be subtracted from .5 and this would equal .0228, which means that 2.28% of individuals would be expected to score better than or equal to 120.

Should a person wish to know the proportion of people who obtained an IQ between 85 and 95, a series of the previous calculations would have to be conducted. In this example, the $z$ scores for 85 and for 95 would have to be first calculated (assuming the same mean and standard deviation as the previous example):

$$\text{For } 85: z = \frac{(X - \mu)}{\sigma} = \frac{(85 - 100)}{10} = -1.5.$$

$$\text{For } 95: z = \frac{(X - \mu)}{\sigma} = \frac{(95 - 100)}{10} = -0.5.$$

The areas under the negative $z$ scores are the same as the areas under the identical positive z-scores because the standard normal distribution is symmetrical about its mean (not all tables present both the positive and negative $z$ scores). In looking up the table for $z = 1.5$ and $z = 0.5$, .4332 and .1915, respectively, are obtained. To determine the proportion of IQ scores between 85 and 95, .1915 has to be subtracted from .4332; .4332 – .1915 = .2417, 24.17% of IQ scores are found between 85 and 95.

For both examples presented here, the proportions are estimates based on mathematical calculations and do not represent actual observations. Therefore, in the previous example, it is estimated that 97.72% of IQ scores are expected to be less than or equal to 120, and it is estimated that 24.17% are expected to be found between 85 and 95, but what would actually be observed might be different.

The normal distribution is also important because of its numerous mathematical properties. Assuming that the data of interest are normally distributed allows researchers to apply different calculations that can only be applied to data that share the characteristics of a normal curve. For instance, many scores such as percentiles, $t$ scores (scores that have been converted to standard scores and subsequently modified such that their mean is 50 and standard deviation is 10), and stanines (scores that have been changed to a value from 1 to 9 depending on their location in the distribution; e.g., a score found in the top 4% of the distribution is given a value of 9, a score found in the middle 20% of the distribution is given a value of 5) are calculated based on the normal distribution. Many statistics rely on the normal distribution as they are based on the assumption that directly observable scores are normally distributed or have a distribution that approximates normality. Some statistics that assume the variables under study are normally distributed include $t, F,$ and $\chi^2$. Furthermore, the normal distribution can be used as an approximation for some other distributions.

To determine whether a given set of data follows a normal distribution, examination of skewness and kurtosis, the Probability-Probability (P-P) plot, or results of normality tests such as the Kolmogorov–Smirnov test, Lilliefors test, and the Shapiro–Wilk test can be conducted. If the data do not reflect a normal distribution, then the researcher has to determine whether a few outliers are influencing the distribution of the data, whether data transformation will be necessary, or

whether nonparametric statistics will be used to analyze the data, for instance.

Many measurements (latent variables) and phenomena are assumed to be normally distributed (and thus can be approximated by the normal distribution). For instance, intelligence, weight, height, abilities, and personality traits can each be said to follow a normal distribution. However, realistically, researchers deal with data that come from populations that do not perfectly follow a normal distribution, or their distributions are not actually known. The Central Limit Theorem (also known as the second fundamental theorem of probability) partly takes care of this problem. One important element of the Central Limit Theorem states that when the sample size is large, the sampling distribution of the sample means will approach the normal curve even if the population distribution is not normal. This allows researchers to be less concerned about whether the population distributions follow a normal distribution or not.

These descriptions and applications apply to the univariate normal distribution (i.e., the normal distribution of a single variable). When two (bivariate normal distribution) or more variables are considered, the multivariate normal distribution is important for examining the relation of those variables and for using multivariate statistics.

Whether many variables are actually normally distributed is a point of debate for many researchers. For instance, the view that certain personality traits are normally distributed can never be observed, as the constructs are not actually measured. Many variables are measured using discrete rather than continuous scales. Furthermore, large sample sizes are not always obtained, thus the normal curve might not actually fit well those data.

## History of the Normal Distribution

The first known documentation of the normal distribution was written by Galileo in the 17th century in his description of random errors found in measurements by astronomers. Abraham de Moivre is credited with its first appearance in his publication of an article in 1733. Pierre Simon de Laplace developed the first general Central Limit Theorem in the early 1800s (an important element in the application of the normal distribution) and described the normal distribution. Carl Friedrich Gauss independently discovered the normal curve and its properties at around the same time as de Laplace and was interested primarily in its application to errors of observation in astronomy. It was consequently extensively used for describing errors. Adolphe Quetelet extended the use of the normal curve beyond errors, believing it could be used to describe phenomena in the social sciences, not just physics. Sir Francis Galton in the late 19th century extended Quetelet's work and applied the normal curve to other psychological measurements.

*Adelheid A. M. Nicol*

***See also*** Central Limit Theorem; Data Cleaning; Multivariate Normal Distribution; Nonparametric Statistics; Normality Assumption; Normalizing Data; Parametric Statistics; Percentile Rank; Sampling Distributions

### Further Readings

Hays, W. L. (1994). *Statistics.* Orlando, FL: Harcourt.

Hopkins, K. D., & Glass, G. V. (1978). *Basic statistics for the behavioral sciences.* Englewood Cliffs, NJ: Prentice Hall.

King, B. M., & Minium, E. M. (2006). *Statistical reasoning in psychology and education.* Hoboken, NJ: Wiley.

Levin, J., & Fox, J. A. (2006). *Elementary statistics in social research.* Boston: Allyn & Bacon.

Lewis, D. G. (1957). The normal distribution of intelligence: A critique. *British Journal of Psychology, 48,* 98–104.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105,* 156–166.

Patel, J. K., & Read, C. B. (1996). *Handbook of the normal distribution.* New York: Marcel Dekker.

Sigler, S. M. (1986). *The history of statistics. The measurement of uncertainty before 1900.* Cambridge, MA: Belknap Press of Harvard University Press.

Snyder, D. M. (1986). On the theoretical derivation of the normal distribution for psychological phenomena. *Psychological Reports, 59,* 399–404.

Thode, H. (2002). *Testing for normality.* New York: Marcel Dekker.

Wilcox, R. R. (1996). *Statistics for the social sciences.* San Diego, CA: Academic Press.

Zimmerman, D. W. (1998). Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions. *Journal of Experimental Education, 67,* 55–68.

## Normality Assumption

The normal distribution (also called the Gaussian distribution: named after Johann Gauss, a German scientist and mathematician who justified the least squares method in 1809) is the most widely used family of statistical distributions on which many statistical tests are based. Many measurements of physical and psychological phenomena can be approximated by the normal distribution and, hence, the widespread utility of the distribution. In many areas of research, a sample is identified on which measurements of particular phenomena are made. These measurements are then statistically tested, via hypothesis testing, to determine whether the observations are different because of chance. Assuming the test is valid, an inference can be made about the population from which the sample is drawn.

Hypothesis testing involves assumptions about the underlying distribution of the sample data. Three key assumptions, in the order of importance, are independence, common variance, and normality. The term *normality assumption* arises when the researcher asserts that the distribution of the data follows a normal distribution. Parametric and nonparametric tests are commonly based on the same assumptions with the exception being nonparametric tests do not require the normality assumption.

Independence refers to the correlation between observations of a sample. For example, if you could order the observations in a sample by time, and observations that are closer together in time are more similar and observations further apart in time are less similar, then we would say the observations are not independent but correlated or dependent on time. If the correlation between observations is positive then the Type I error is inflated (Type I error level is the probability of rejecting the null hypothesis when it is true and is traditionally defined by alpha and set at .05). If the correlation is negative, then Type I error is deflated. Even modest levels of correlation can have substantial impacts on the Type I error level (for a correlation of .2 the alpha is .11, whereas for a correlation of .5, the alpha level is .26). Independence of observations is difficult to assess. With no formal statistical tests widely in use, knowledge of the substantive area is paramount and a through understanding of how the data were generated is required for valid statistical analysis and interpretation to be undertaken.

Common variance (often referred to as homogeneity of variance) refers to the concept that the variance of all samples drawn has similar variability. For example, if you were testing the difference in height between two samples of people, one from Town A and the other from Town B, the test assumes that the variance of height in Town A is similar to that of Town B. In 1953, G. E. P. Box demonstrated that for even modest sample sizes, most tests are robust to this assumption, and differences of up to 3-fold in variance do not greatly affect the Type I error level. Many statistical tests are available to ascertain whether the variances are equal among different samples (including the Bartlett–Kendall test, Levene's test, and the Brown–Forsythe test). These tests for the homogeneity of variance are sensitive to normality departures, and as such they might indicate that the common variance assumption does not hold, although the validity of the test is not in question.

Although it is the least important of the assumptions when considering hypothesis testing, the normality assumption should not be ignored. Many statistical tests and methods employed in research require that a variable or variables be normally distributed. Aspects of normality include the following: (a) that the possible values of the quantity being studied can vary from negative infinity to positive infinity; (b) that there will be symmetry in the data, in that observed values will fall with equal probability above and below the true population mean value, as a process of unbiased, random variability; and (c) the width or spread of the distribution of observed values around the true mean will be determined by the standard deviation of the distribution (with predictable percentages of observed values falling bounds defined by multiples of the true distribution standard deviation). Real-world data might behave differently from the normal distribution for several reasons. Many physiological and behavioral variables cannot truly have infinite range, and distributions might be truncated at true zero values. Skewness (asymmetry), or greater spread in observed values than predicted by the standard deviation, might also result in situations where the distribution of observed

values is not determined exclusively by random variability, and it also might be a result of unidentified systematic influences (or unmeasured predictors of the outcome).

The statistical tests assume that the data follow a normal distribution to preserve the tests' validity. When undertaking regression models, the normality assumption applies to the error term of the model (often called the residuals) and not the original data and, hence, it is often misunderstood in this context. It should be noted that the normality assumption is sufficient, but not necessary, for the validity of many hypothesis tests. The remainder of this entry focuses on the assessment of normality and the transformation of data that are not normally distributed.

## Assessing Normality

A researcher can assess for the normality of variables in several ways. To say a variable is normally distributed indicates that the distribution of observations for that variable follows the normal distribution. So in essence, if you examined the distribution graphically, it would look similar to the typical bell-shaped normal cure. A histogram, a box-and-whisker plot, or a normal quartile plot (often called a Q-Q plot) can be created to inspect the normality of the data visually. With many data analysis packages, a histogram can be requested with the normal distribution superimposed to aid in this assessment. Other standard measures of distribution exist and include skewness and kurtosis. Skewness refers to the symmetry of the distribution, in which right-skewed distributions have a long tail pointing to the right and left-skewed distributions have a long tail pointing to the left. Kurtosis refers to peakedness of the distribution.

A box-and-whisker plot is created with five numeric summaries of the variables including the minimum value, the lower quartile, the median, the upper quartile, and the maximum value. The box is formed by the lower and upper quartile bisected by the median. Whiskers are formed on the box plot by drawing a line from the lowest edge of the box (lower quartile) to the minimum value and the highest edge of the box (upper quartile) to the maximum value. If the variable has a normal distribution, the box will be bisected in the middle by the median, and both whiskers will be of equal length.

A normal quartile plot compares the spacing of the data with that of the normal distribution. If the data being examined are approximately normal, then more observations should be clustered around the mean and only a few observations should exist in each of the tails. The vertical axis of the plot displays the actual data whereas the horizontal axis displays the quartiles from the normal distribution (expected $z$ scores). If the data are normally distributed, the resulting plot will form a straight line with a slope of 1. If the line demonstrates an upward bending curve, the data are right skewed, whereas if the line demonstrates a downward bending curve, the data are left skewed. If the line has an S-shape, it indicates that the data are kurtotic.

Several common statistical tests were designed to assess normality. These would include, but are not limited to, the Kolmogorov–Smirnov, the Shapiro–Wilk, the Anderson–Darling, and the Lilliefor's test. In each case, the test calculates a test statistic under the null hypothesis that the sample is drawn from a normal distribution. If the associated $p$ value for the test statistic is greater than the selected alpha level then one does not reject the null hypothesis that the data were drawn from a normal distribution. Some tests can be modified to test samples against other statistical distributions. The Shapiro–Wilk and the Anderson–Darling tests have been noted to perform better with small sample sizes. With all the tests, small deviations from normality can lead to a rejection of the null hypothesis and therefore should be used and interpreted with caution.

When a violation of the normality assumption is observed, it might be a sign that a better statistical model can be found. So, exploring why the assumption is violated might be fruitful. Nonnormality of the error term might indicate that the resulting error is greater than expected under the assumption of true random variability and (especially when the distribution of the data are asymmetrical) might suggest that the observations come from more than one "true" underlying population. Additional variables could be added to the model (or the study) to predict systematically observed values not yet in the model, thereby moving more information to the linear predictor. Similarly, nonnormality might reflect that variables in the model

are incorrectly specified (such as assuming there is a linear association between a continuous predictor variable and the outcome).

Research is often concerned with more than one variable, and with regression analysis or statistical modeling, the assumption is that the combination of variables under study follows a multivariate normal distribution. There are no direct tests for multivariate normality, and as such, each variable under consideration is considered individually for normality. If all the variables under study are normally distributed, then another assumption is made that the variables combined are multivariate normal. Although this assumption is made, it is not always the case that variables are normal individually and collectively. Note that when assessing normality in a regression modeling situation, the assessment of normality should be undertaken with the error term (residuals).

### *Note of Caution: Small Sample Sizes*

When assessing normality with small sample sizes (samples with less than approximately 50 observations), caution should be exercised. Both the visual aids (histogram, box-and-whisker plot, and normal quartile plot) and the statistical tests (Kolmogorov–Smirnov, Shapiro–Wilk, Anderson–Darling, and Lilliefor's test) can provide misleading results. Departures from normality are difficult to detect with small sample sizes, largely because of the power of the test. The power of the statistical tests decreases as the significance level is decreased (as the statistical test is made more stringent) and increases as the sample size increases. So, with small sample sizes, the statistical tests will nearly always indicate acceptance of the null hypothesis even though departures from normality could be large. Likewise, with large sample sizes, the statistical tests become powerful, and often minor inconsequential departures from normality would lead the researcher to reject the null hypothesis.

## What to Do If Data Are Not Normally Distributed

If, after assessment, the data are not normally distributed, a transformation of the non-normal variables might improve normality. If after transformation the variable meets the normality assumption, the transformed variable can be substituted in the analysis. Interpretation of a transformed variable in an analysis needs to be undertaken with caution as the scale of the variable will be related to the transformation and not the original units.

Based on Frederick Mosteller and John Tukey's Ladders of Power, if a researcher needs to remove right skewness from the data, then he or she moves "down" the ladder of power by applying a transformation smaller than 1, such as the square root, cube root, logarithm, or reciprocal. If the researcher needs to remove left skewness from the data, then he or she moves "up" the ladder of power by applying a transformation larger than 1 such as squaring or cubing.

Many analysts have tried other means to avoid non-normality of the error term including categorizing the variable, truncating or eliminating extreme values from the distribution of the original variable, or restricting the study or experiment to observations within a narrower range of the original measure (where the residuals observed might form a "normal" pattern). None of these are ideal as they might affect the measurement properties of the original variable and create problems with bias of estimates of interest and/or loss of statistical power in the analysis.

*Jason D. Pole and Susan J. Bondy*

***See also*** Central Limit Theorem; Homogeneity of Variance; Law of Large Numbers; Normal Distribution; Type I Error; Type II Error; Variance

**Further Readings**

Box, G. E. P. (1953). Non-normality and tests on variances. *Biometrika,* 40, 318–335.

Holgersson, H. E. T. (2006). A graphical method for assessing multivariate normality. *Computational Statistics,* 21, 141–149.

Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Annual Review of Public Health,* 23, 151–169.

# Normalizing Data

Researchers often want to compare scores or sets of scores obtained on different scales. For example, how do we compare a score of 85 in a cooking contest with a score of 100 on an IQ test? To do so, we need to "eliminate" the unit of measurement; this operation means to *normalize* the data. There are two main types of normalization. The first type of normalization originates from linear algebra and treats the data as a *vector* in a multidimensional space. In this context, to normalize the data is to transform the data vector into a new vector whose *norm* (i.e., length) is equal to one. The second type of normalization originates from statistics and eliminates the unit of measurement by transforming the data into new scores with a mean of 0 and a standard deviation of 1. These transformed scores are known as $z$ scores.

## Normalization to a Norm of One

### *The Norm of a Vector*

In linear algebra, the *norm* of a vector measures its *length* which is equal to the Euclidean distance of the endpoint of this vector to the origin of the vector space. This quantity is computed (from the Pythagorean theorem) as the square root of the sum of the squared elements of the vector. For example, consider the following data vector denoted $\mathbf{y}$:

$$\mathbf{y} = \begin{bmatrix} 35 \\ 36 \\ 46 \\ 68 \\ 70 \end{bmatrix}. \tag{1}$$

The norm of vector $\mathbf{y}$ is denoted $||\mathbf{y}||$ and is computed as

$$\begin{aligned} \|\mathbf{y}\| &= \sqrt{35^2 + 36^2 + 46^2 + 68^2 + 70^2} \\ &= \sqrt{14,161} = 119. \end{aligned} \tag{2}$$

### *Normalizing With the Norm*

To normalize $\mathbf{y}$, we divide each element by $||\mathbf{y}|| = 119$. The normalized vector, denoted $\tilde{\mathbf{y}}$, is equal to

$$\tilde{\mathbf{y}} = \begin{bmatrix} \frac{35}{119} \\ \frac{36}{119} \\ \frac{46}{119} \\ \frac{68}{119} \\ \frac{70}{119} \end{bmatrix} = \begin{bmatrix} 0.2941 \\ 0.3025 \\ 0.3866 \\ 0.5714 \\ 0.5882 \end{bmatrix}. \tag{3}$$

The norm of vector $\tilde{\mathbf{y}}$ is now equal to one:

$$\begin{aligned} &\|\tilde{\mathbf{y}}\| = \\ &\sqrt{0.2941^2 + 0.3025^2 + 0.3866^2 + 0.5714^2 + 0.5882^2} \\ &= \sqrt{1} = 1. \end{aligned} \tag{4}$$

## Normalization Using Centering and Standard Deviation: $z$ Scores

### *The Standard Deviation of a Set of Scores*

Recall that the standard deviation of a set of scores expresses the dispersion of the scores around their mean. A set of $N$ scores, each denoted $Y_n$, whose mean is equal to $M$, has a standard deviation denoted $\hat{S}$ which is computed as

$$\hat{S} = \sqrt{\frac{\sum (Y_N - M)^2}{N - 1}}. \tag{5}$$

For example, the scores from vector $\mathbf{y}$ (see Equation 4) have a mean of 51 and a standard deviation of

$$\begin{aligned} \hat{S} &= \sqrt{\frac{(35-51)^2 + (36-51)^2 + (46-51)^2 + (68-51)^2 + (70-51)^2}{5-1}} \\ &= \frac{1}{2}\sqrt{(-16)^2 + -15^2 + (-5)^2 + 17^2 + 19^2} \\ &= 17. \end{aligned} \tag{6}$$

### z Scores: Normalizing With the Standard Deviation

To normalize a set of scores using the standard deviation, we divide each score by the standard deviation of this set of scores. In this context, we almost always subtract the mean of the scores from each score prior to dividing by the standard deviation. This normalization is known as $z$ scores. Formally, a set of $N$ scores each denoted $Y_n$ and whose mean is equal to $M$ and whose standard deviation is equal to $\hat{S}$ is transformed in $z$ scores as

$$z_n = \frac{Y_n - M}{\hat{S}}. \quad (7)$$

With elementary algebraic manipulations, it can be shown that a set of $z$ scores has a mean equal of zero and a standard deviation of one. Therefore, $z$ scores constitute a unit-free measure that can be used to compare observations measured with different units.

#### Example

For example, the scores from vector $\mathbf{y}$ (see Equation 1) have a mean of 51 and a standard deviation of 17. These scores can be transformed into the vector $\mathbf{z}$ of $z$ scores as

$$\mathbf{z} = \begin{bmatrix} \frac{35-51}{17} \\ \frac{36-51}{17} \\ \frac{46-51}{17} \\ \frac{68-51}{17} \\ \frac{70-51}{17} \end{bmatrix} = \begin{bmatrix} -\frac{16}{17} \\ -\frac{15}{17} \\ -\frac{5}{17} \\ \frac{17}{17} \\ \frac{19}{17} \end{bmatrix} = \begin{bmatrix} -0.9412 \\ -0.8824 \\ -0.2941 \\ 1.0000 \\ 1.1176 \end{bmatrix}. \quad (8)$$

The mean of vector $\mathbf{z}$ is now equal to zero, and its standard deviation is equal to one.

*Hervé Abdi and Lynne J. Williams*

*See also* Mean; Normal Distribution; Standard Deviation; Standardization; Standardized Score; Variance; $z$ Score

#### Further Readings

Abdi, H., Edelman, B., Valentin, D., & Dowling, W. J. (2009). *Experimental design and analysis for psychology.* Oxford, UK: Oxford University Press.

Date, C. J., Darwen, H., & Lorentzos, N. A. (2003). *Temporal data and the relational model.* San Fancisco: Morgan Kauffman.

Myatt, G. J., & Johnson, W. P. (2009). *Making Sense of Data II: A Practical Guide to Data Visualization, Advanced Data Mining Methods, and Applications.* Hoboken, NJ: Wiley.

## Nuisance Variable

A nuisance variable is an unwanted variable that is typically correlated with the hypothesized independent variable within an experimental study but is typically of no interest to the researcher. It might be a characteristic of the participants under study or any unintended influence on an experimental manipulation. A nuisance variable causes greater variability in the distribution of a sample's scores and affects all the groups measured within a given sample (e.g., treatment and control).

Whereas the distribution of scores changes because of the nuisance variable, the location (or the measure of central tendency) of the distribution remains the same, even when a nuisance variable is present. Figure 1 is an example using participants' scores on a measure that assesses accuracy in responding to simple math problems. Note that the mean is 5.00 for each of the two samples, yet the distributions are actually different. The range for the group without the nuisance variable is 4, with a standard deviation of 1.19. However, the range

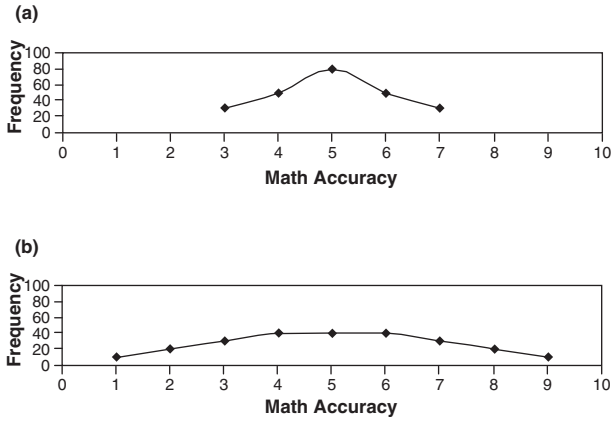



**Figure 1** Frequency Distributions With and Without Nuisance Variable

*Notes:* (a) Without nuisance variable ($N = 240$). (b) With nuisance variable ($N = 240$).

for the group with the nuisance variable is 8, with a standard deviation of 2.05.

For the first distribution, participants' math accuracy scores are relatively similar and cluster around the mean of the distribution. There are fewer very high scores and fewer very low scores when compared with the distribution in which the nuisance variable is operating. Within the distribution with the nuisance variable, a wider spread is observed, and there are fewer scores at the mean than in the distribution without the nuisance variable.

In an experimental study, a nuisance variable affects within-group differences for both the treatment group and the control group. When a nuisance variable is present, the spread of scores for each group increases, which makes it more difficult to observe effects that might be attributed to the independent variable (i.e., the treatment effect). When there is greater spread within the distributions of the treatment group and the control group, there is more overlap between the two. This makes the differences between the two groups less clear and distinct.

Figure 2 is an example using treatment and control groups, in which the independent variable is an extracurricular math tutoring program and is received only by the treatment group. Both samples are tested after the administration of the math tutoring program on measures of math accuracy. In this case, a nuisance variable might be the participant's level of anxiety, but it could just as easily be an external characteristic of the experiment, such as the amount of superfluous noise in the room where the measure of math accuracy is being administered. In considering the effects of the nuisance variable, the distributions of the two groups within the sample might look similar to that in Figure 2.

In the distribution without the nuisance variable, the variation in participants' anxiety is reduced or eliminated, and it is clear that the differences in participants' observed math scores are, more than likely, caused by the manipulation of the independent variable only (i.e., the administration of a math tutorial to the treatment group but not the control group).

In the distribution with the nuisance variable, if there is greater variation in the amount of anxiety experienced by participants in both the treatment group and the control group, and if the participant's

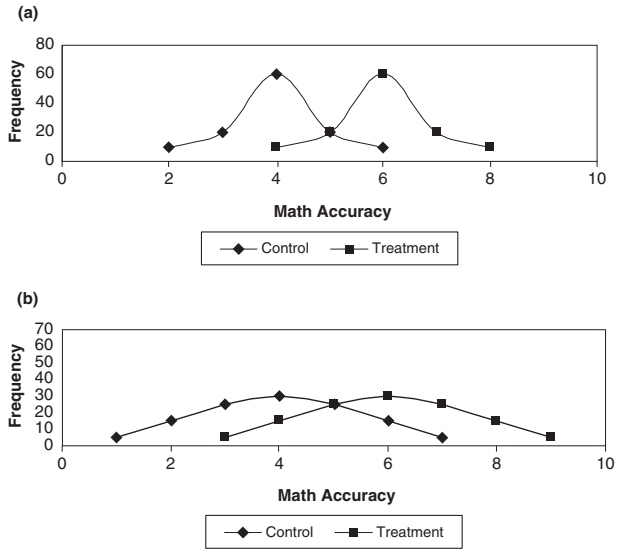



**Figure 2** Frequency Distributions of Treatment and Control Groups With and Without Nuisance Variable

*Notes:* (a) Without nuisance variable. (b) With nuisance variable.

anxiety influences his or her math accuracy, then greater variation will be observed in the distributions of participants' scores on the measure of math accuracy. For example, someone who performs poorly when feeling anxious might have scored a 4 on math accuracy if this nuisance variable had been removed. However, now that he or she is feeling anxious, his or her score might drop to a 2. Alternatively, others might perform better when experiencing anxiety. In this case, a participant might have originally scored a 6 if he or she felt no anxiety but actually scored an 8 because of the anxiety. In this case, with the nuisance variable present, it becomes more difficult to detect whether statistically significant differences exist between the two groups in question because of the greater variance within the groups.

Researchers aim to control the influences of nuisance variables methodologically and/or statistically, so that differences in the dependent variable might be attributed more clearly to the manipulation of the independent variable. In exercising methodological control, a researcher might choose only those participants who do not experience anxiety when taking tests of this nature. This would eliminate any variation in anxiety and, in turn, any influence that anxiety might have on

participants' math accuracy scores. In exercising statistical control, a researcher might employ regression techniques to control for any variation caused by a nuisance variable. However, in this case, it becomes important to specify and measure potential nuisance variables before and during the experiment. In the example used previously, participants could be given a measure of anxiety along with the measure of math accuracy. Multiple regression models might then be used to statistically control for the influence of anxiety on math accuracy scores alongside any experimental treatment that might be administered.

The term *nuisance variable* is often used alongside the terms *extraneous* and *confounding* variable. Whereas an extraneous variable influences differences observed between groups, a nuisance variable influences differences observed within groups. By eliminating the effects of nuisance variables, the tests of the null hypothesis become more powerful in uncovering group differences.

*Cynthia R. Davis*

***See also*** Confounding; Control Variables; Statistical Control

**Further Readings**

Breaugh, J. A. (2006). Rethinking the control of nuisance variables in theory testing. *Journal of Business and Psychology, 20,* 429–443.

Meehl, P. (1970). Nuisance variables and the ex post facto design. In M. Radner & S. Winokur (Eds.), *Minnesota studies in the philosophy of science: Vol. IV. Analyses of theories and methods of physics and psychology* (pp. 373–402). Minneapolis: University of Minnesota Press.

# Null Hypothesis

In many sciences, including ecology, medicine, and psychology, null hypothesis significance testing (NHST) is the primary means by which the numbers comprising the data from some experiment are translated into conclusions about the question(s) that the experiment was designed to address. This entry first provides a brief description of NHST, and within the context of NHST, it defines the most common incarnation of a null hypothesis. Second, this entry sketches other less common forms of a null hypothesis. Third, this entry articulates several problems with using null hypothesis-based data analysis procedures.

## Null Hypothesis Significance Testing and the Null Hypothesis

Most experiments entail measuring the effect(s) of some number of independent variables on some dependent variable.

### *An Example Experiment*

In the simplest sort of experimental design, one measures the effect of a single independent variable, such as the amount of information held in short-term memory on a single dependent variable and the reaction time to scan through this information. To pick a somewhat arbitrary example from cognitive psychology, consider what is known as a *Sternberg experiment,* in which a short sequence of memory digits (e.g., "34291") is read to an observer who must then decide whether a single, subsequently presented test digit was part of the sequence. Thus for instance, given the memory digits above, the correct answer would be "yes" for a test digit of "2" but "no" for a test digit of "8." The independent variable of "amount of information held in short-term memory" can be implemented by varying *set size,* which is the number of memory digits presented: In different conditions, the set size might be, say, 1, 3, 5 (as in the example), or 8 presented memory digits. The number of different set sizes (here 4) is more generally referred to as the number of *levels* of the independent variable. The dependent variable is the reaction time measured from the appearance of the test digit to the observer's response. Of interest in general is the degree to which the magnitude of the dependent variable (here, reaction time) depends on the level of the independent variable (here set size).

### *Sample and Population Means*

Typically, the principal dependent variable takes the form of a *mean.* In this example, the mean reaction time for a given set size could be computed across observers. Such a computed mean is called a *sample mean,* referring to its having been computed across an observed sample of numbers.

A sample mean is construed as an estimate of a corresponding *population mean,* which is what the mean value of the dependent variable would be if all observers in the relevant population were to participate in a given condition of the experiment. Generally, conclusions from experiments are meant to apply to population means. Therefore, the measured sample means are only interesting insofar as they are estimates of the corresponding population means.

Notationally, the sample means are referred to as the $M_j$s, whereas the population means are referred to as the $\mu_j$s. For both sample and population means, the subscript "$j$" indexes the level of the independent variable; thus, in our example, $M_2$ would refer to the observed mean reaction time of the second set-size level (i.e., set size = 3) and likewise, $\mu_2$ would refer to the corresponding, unobservable population mean reaction time corresponding to set size = 3.

### *Two Competing Hypotheses*

NHST entails establishing and evaluating two mutually exclusive and exhaustive hypotheses about the relation between the independent variable and the dependent variable. Usually, and in its simplest form, the null hypothesis (abbreviated $H_0$) is that the independent variable has *no effect* on the dependent variable, whereas the alternative hypothesis (abbreviated $H_1$) is that the independent variable has *some* effect on the dependent variable. Note an important asymmetry between a null hypothesis and an alternative hypothesis: A null hypothesis is an *exact* hypothesis, whereas an alternative hypothesis is an *inexact* hypothesis. By this it is meant that a null hypothesis can be correct in only one way, viz, the $\mu_j$s are all equal to one another, whereas there are an infinite number of ways in which the $\mu_j$s can be different from one another (i.e., an infinite number of ways in which an alternative hypothesis can be true).

### *Decisions Based on Data*

Having established a null and an alternative hypothesis that are mutually exclusive and exhaustive, the experimental data are used to—roughly speaking—decide between them. The technical manner by which one makes such a decision is beyond the scope of this entry, but two remarks about the process are appropriate here.

1. A major ingredient in the decision is the variability of the $M_j$s. To the degree that the $M_j$s are close to one another, evidence ensues for possible equality of the $\mu_j$s and, ipso facto, validity of the null hypothesis. Conversely, to the degree that the $M_j$s differ from one another, evidence ensues for associated differences among the $\mu_j$s and, ipso facto, validity of the alternative hypothesis.

2. The asymmetry between the null hypothesis (which is exact) and the alternative hypothesis (which is inexact) sketched previously implies an associated asymmetry in conclusions about their validity. If the $M_j$s differ sufficiently, one "rejects the null hypothesis" in favor of accepting the alternative hypothesis. However, if the $M_j$s do not differ sufficiently, one does not "accept the null hypothesis" but rather one "fails to reject the null hypothesis." The reason for the awkward, but logically necessary, wording of the latter conclusion is that, because the alternative hypothesis is inexact, one cannot generally distinguish a genuinely true null hypothesis on the one hand from an alternative hypothesis entailing small differences among the $\mu_j$s on the other hand.

### *Multifactor Designs: Multiple Null Hypothesis–Alternative Hypothesis Pairings*

So far, this entry has described a simple design in which the effect of a single independent variable on a single dependent variable is examined. Many, if not most experiments, use multiple independent variables and are known as *multifactor designs* ("factor" and "independent variable" are synonymous). Continuing with the example experiment, imagine that in addition to measuring the effects of set size on reaction time in a Sternberg task, one also wanted to measure simultaneously the effects on reaction time of the test digit's visual contrast (informally, the degree to which the test digit stands out against the background). One might then *factorially combine* the four levels of set size (now called "factor 1") with, say, two levels, "high contrast" and "low contrast," of test-digit contrast (now called "factor 2"). Combining the four set-size levels with the two test-digit contrast levels would yield $4 \times 2 = 8$ separate conditions.

Typically, three independent NHST procedures would then be carried out, entailing three null hypothesis–alternative hypothesis pairings. They are as follows:

**1. For the set size main effect:**

$H_0$ : Averaged over the two test-digit contrasts, there is no set-size effect

$H_1$ : Averaged over the two test-digit contrasts, there is a set-size effect

**2. For the test-digit contrast main effect:**

$H_0$ : Averaged over the four set sizes, there is no test-digit contrast effect

$H_1$ : Averaged over the four set sizes, there is a test-digit contrast effect

**3. For set-size by test-digit contrast interaction:**

Two independent variables are said to *interact* if the effect of one independent variable depends on the level of the other independent variable. As with the main effects, interaction effects are immediately identifiable with respect to the $M_j$s; however, again as with main effects, the goal is to decide whether interaction effects exist with respect to the corresponding $\mu_j$s. As with the main effects, NHST involves pitting a null hypothesis against an associated alternative hypothesis.

$H_0$ : With respect to the $\mu_j$s, set size and test-digit contrast do not interact.

$H_1$ : With respect to the $\mu_j$s, set size and test-digit contrast do interact.

The logic of carrying out NHST with respect to interactions is the same as the logic of carrying out NHST with respect to main effects. In particular, with interactions as with main effects, one can reject a null hypothesis of no interaction, but one cannot accept a null hypothesis of no interaction.

## Non-Zero-Effect Null Hypotheses

The null hypotheses described above imply "no effect" of one sort or another—either no main effect of some independent variable or no interaction between two independent variables. This kind of no-effect null hypothesis is by far the most common null hypothesis to be found in the literature. Technically however, a null hypothesis can be any *exact hypothesis;* that is, the null hypothesis of "all $\mu_j$s are equal to one another" is but one special case of what a null hypothesis can be.

To illustrate another form, let us continue with the first, simpler Sternberg-task example (set size is the only independent variable), but imagine that prior research justifies the assumption that the relation between set size and reaction time is linear. Suppose also that research with digits has yielded the conclusion that reaction time increases by 35 ms for every additional digit held in short-term memory; that is, if reaction time were plotted against set size, the resulting function would be linear with a slope of 35 ms.

Now, let us imagine that the Sternberg experiment is done with words rather than digits. One could establish the null hypothesis that "short-term memory processing proceeds at the same rate with words as it does with digits" (i.e., that the slope of the reaction time versus set-size function would be 35 ms for words just as it is known to be with digits). The alternative hypothesis would then be "for words, the function's slope is anything other than 35 ms." Again, the fundamental distinction between a null and alternative hypothesis is that the null hypothesis is exact (35 ms/digit), whereas the alternative hypothesis is inexact (anything else). This distinction would again drive the asymmetry between conclusions, which were articulated previously: a particular pattern of empirical results could logically allow "rejection of the null hypothesis"; that is, "acceptance of the alternative hypothesis" but not "acceptance of the null hypothesis."

## Problems With Null Hypothesis Significance Testing

No description of NHST in general, or a null hypothesis in particular, is complete without at least a brief account of the serious problems that accrue when NHST is the sole statistical technique used for making inferences about the $\mu$s from the $M_j$s. Briefly, three major problems involving a null hypothesis as the centerpiece of data analysis are discussed below.

### *A Null Hypothesis Cannot Be Literally True*

In most sciences, it is almost a self-evident truth that any independent variable must have some effect, even if small, on any dependent variable. This is certainly true in psychology. In the Sternberg task, to illustrate, it is simply implausible that set size would have literally *zero* effect on reaction time (i.e., that is, that the $\mu_j$s corresponding to the different set sizes would be *identical* to an infinite number of decimal places). Therefore, rejecting a null hypothesis—which, as noted, is the only strong conclusion that is possible within the context of NHST—tells the investigator nothing that the investigator should have been able to realize was true beforehand. Most investigators do not recognize this, but that does not prevent it from being so.

### *Human Nature Makes Acceptance of a Null Hypothesis Almost Irresistible*

Earlier, this entry detailed why it is logically forbidden to accept a null hypothesis. However, human nature dictates that people do not like to make weak yet complicated conclusions such as "We fail to reject the null hypothesis." Scientific investigators, generally being humans, are not exceptions. Instead, a "fail to reject" decision, which is dutifully made in an article's results section, often morphs into "the null hypothesis is true" in the article's discussion and conclusions sections. This kind of sloppiness, although understandable, has led to no end of confusion and general scientific mischief within numerous disciplines.

### *Null Hypothesis Significance Testing Emphasizes Barren, Dichotomous Conclusions*

Earlier, this entry described that the pattern of population means—the relations among the unobservable $\mu_j$s—is of primary interest in most scientific experiments and that the observable $M_j$s are estimates of the $\mu_j$s. Accordingly, it should be of great interest to assess how *good* are the $M_j$s as estimates of the $\mu_j$s. If, to use an extreme example, the $M_j$s were perfect estimates of the $\mu_j$s there would be no need for statistical analysis: The answers to any question about the $\mu_j$s would be immediately available from the data. To the degree that the estimates are less good, one must exercise concomitant caution in using the $M_j$s to make inferences about the $\mu_j$s.

None of this is relevant within the process of NHST, which does not in any way emphasize the degree to which the $M_j$s are good estimates of the $\mu_j$s. In its typical form, NHST allows only a limited assessment of the nature of the $\mu_j$s: Are they all equal or not? Typically, the "no" or "not necessarily no" conclusion that emerges from this process is insufficient to evaluate the totality of what the data might potentially reveal about the nature of the $\mu_j$s.

An alternative that is gradually emerging within several NHST-heavy sciences—an alternative that is common in the natural sciences—is the use of *confidence intervals* that assess directly how good is a $M_j$ as an estimate of the corresponding $\mu_j$. Briefly, a confidence interval is an interval constructed around a sample mean that, with some pre-specified probability (typically 95%), includes the corresponding population mean. A glance at a set of plotted $M_j$s with associated plotted confidence intervals provides immediate and intuitive information about (a) the most likely pattern of the $\mu_j$s and (b) the reliability of the pattern of $M_j$s as an estimate of the pattern of $\mu_j$s. This in turn provides immediate and intuitive information both about the relatively uninteresting question of whether some null hypothesis is true and about the much more interesting questions of what the pattern of $\mu_j$s actually is and how much belief can be placed in it based on the data at hand.

*Geoffrey R. Loftus*

***See also*** Confidence Intervals; Hypothesis; Research Hypothesis; Research Question

#### Further Readings

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods, 10,* 389–396.

Cumming, G., Williams, J., & Fidler, F. (2004). Replication, and researchers' understanding of confidence intervals and standard error bars. *Understanding Statistics, 3,* 299–311.

Fidler, F., Burgman, M., Cumming, G., Buttrose, R., & Thomason, N. (2006). Impact of criticism of null hypothesis significance testing on statistical reporting practices in conservation biology. *Conservation Biology, 20,* 1539–1544.

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research, 7,* 1–20.
Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.
Lecoutre, M. P., Poitevineau, J., & Lecoutre, B. (2003). Even statisticians are not immune to misinterpretations of null hypothesis significance tests. *International Journal of Psychology, 38,* 37–45.
Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science, 5,* 161–171.
Rosenthal, R., & Gaito, J. (1963). The interpretation of levels of significance by psychological researchers. *Journal of Psychology, 55,* 33–38.

# Nuremberg Code

The term *Nuremberg Code* refers to the set of standards for conducting research with human subjects that was developed in 1947 at the end of World War II, in the trial of 23 Nazi doctors and scientists in Nuremberg, Germany, for war crimes that included medical experiments on persons designated as non-German nationals. The trial of individual Nazi leaders by the Nuremberg War Crimes Tribunal, the supranational institution charged with determining justice in the transition to democracy, set a vital precedent for international jurisprudence.

The *Nuremberg Code* was designed to protect the autonomy and rights of human subjects in medical research, as compared with the Hippocratic Oath applied in the therapeutic, paternalistic patient-physician relationship. It is recognized as initiating the modern international human rights movement during social construction of ethical codes, with the Universal Declaration of Human Rights in 1954. Human subjects abuse by Nazi physicians occurred despite German guidelines for protection in experimentation, as noted by Michael Branigan and Judith Boss. Because the international use of prisoners in research had grown during World War II, the Code required that children, prisoners, and patients in mental institutions were not to be used as subjects in experiments. However, it was reinterpreted to expand medical research in the Declaration of Helsinki by the World Medical Association in 1964.

The legal judgment in the Nuremberg trial by a panel of American and European physicians and scientists contained 10 moral, ethical, and legal requirements to guide researchers in experiments with human subjects. These requirements are as follows: (1) voluntary informed consent based on legal capacity and without coercion is essential; (2) research should be designed to produce results for the good of society that are not obtainable by other means; (3) human subjects research should be based on prior animal research; (4) physical and mental suffering must be avoided; (5) no research should be conducted for which death or disabling injury is anticipated; (6) risks should be justified by anticipated humanitarian benefits; (7) precautions and facilities should be provided to protect research subjects against potential injury, disability, or death; (8) research should only be conducted by qualified scientists; (9) the subject should be able to end the study during the research; (10) the scientist should be able to end the research at any stage if potential for injury, disability, or death of the subject is recognized.

## Impact on Human Subjects Research

At the time the *Nuremberg Code* was formulated, many viewed it as created in response to Nazi medical experimentation and without legal authority in the United States and Europe. Some American scientists considered the guidelines implicit in their human subjects research, applying to nontherapeutic research in wartime. The informed consent requirement was later incorporated into biomedical research, and physicians continued to be guided by the Hippocratic Oath for clinical research.

Reinterpretation of the *Nuremberg Code* in the Declaration of Helsinki for medical research modified requirements for informed consent and subject recruitment, particularly in pediatrics, psychiatry, and research with prisoners. Therapeutic research was distinguished from nontherapeutic research, and therapeutic privilege was legitimated in the patient-physician relationship.

However, social and biomedical change, as well as the roles of scientists and ethicists in research

and technology, led to the creation of an explicitly subjects-centered approach to human rights. This has been incorporated into research in medicine and public health, and in behavioral and social sciences.

## Biomedical, Behavioral, and Community Research

With the enactment of federal civil and patient rights legislation, the erosion of public trust, and the extensive criticism of ethical violations and discrimination in the Tuskegee syphilis experiments by the United States Public Health Service (1932–1971), biomedical and behavioral research with human subjects became regulated by academic and hospital-based institutional review boards. The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research was established in the United States in 1974, with requirements for review boards in institutions supported by the Department of Health, Education and Welfare.

In 1979, the Belmont Report set four ethical principles for human subjects research: (1) beneficence and nonmalfeasance, to maximize benefits and minimize risk of harm; (2) respect for autonomy in decision-making and protection of those with limited autonomy; (3) justice, for fair treatment; and (4) equitable distribution of benefits and risks. The Council for International Organizations of Medical Sciences and the World Health Organization formulated *International Ethical Guidelines for Biomedical Research Involving Human Subjects* for research ethics committees in 1982. Yet in 1987, the United States Supreme Court refused to endorse the *Nuremberg Code* as binding on all research, and it was not until 1997 that national security research had to be based on informed consent.

Institutional review boards (IRBs) were established to monitor informed consent and avoid risk and exploitation of vulnerable populations. Research organizations and Veterans Administration facilities might be accredited by the Association for Accreditation of Human Research Protection Programs. Although IRBs originated for biomedical and clinical research, their purview extends to behavioral and social sciences. Nancy Shore and colleagues note that public health emphasis on community-based participatory research is shifting focus from protection of individual subjects to ethical relationships with community members and organizations as partners.

Yet global clinical drug trials by pharmaceutical companies and researchers' efforts to limit restrictions on placebo-controlled trials have contributed to the substitution of "Good Clinical Practice Rules" for the Declaration of Helsinki by the U.S. Food and Drug Administration in 2004. The development of rules by regulators and drug industry trade groups and the approval by untrained local ethics committees in developing countries could diminish voluntary consent and benefits for research subjects.

Subsequent change in application of the *Nuremberg Code* has occurred with the use of prisoners in clinical drug trials, particularly for HIV drugs, according to Branigan and Boss. This practice is illegal for researchers who receive federal support but is legal in some states. The inclusion of prisoners and recruitment of subjects from ethnic or racial minority groups for clinical trials might offer them potential benefits, although it must be balanced against risk and need for justice.

*Sue Gena Lurie*

***See also*** Declaration of Helsinki

### Further Readings

Beauchamp, D., & Steinbock, B. (1999). *New ethics for the public's health*. Oxford, UK: Oxford University Press.

Branigan, M., & Boss, J. (2001). Human and animal experimentation. In *Healthcare ethics in a diverse society*. Mountain View, CA: Mayfield.

Cohen, J., Bankert, E., & Cooper, J. (2006). History and ethics. *CITI course in the protection of human research subjects*. Retrieved December 6, 2006, from http://www.citiprogram.org/members/courseandexam/moduletext

Elster, J. (2004). *Closing the books: Transitional justice in historical perspective*. New York: Cambridge University Press.

Farmer, P. (2005). Rethinking health and human rights. In *Pathologies of power*. Berkeley: University of California.

Levine, R. (1981). The Nuremberg Code. In *Ethics and the regulation of clinical research*. Baltimore: Urban & Schwarzenberg.

Levine, R. (1996). The Institutional Review Board. In S. S. Coughlin & T. L. Beauchamp (Eds.), *Ethics and epidemiology*. New York: Oxford University.
Lifton, R. (2000). *The Nazi doctors*. New York: Basic Books.
Morrison, E. (2008). *Health care ethics: Critical issues for the 21st century*. Sudbuy, MA: Jones & Bartlett.
National commission for the protection of human subjects of biomedical and behavioral research. (1979). *The Belmont Report: Ethical principles and guidelines for the protection of human subjects of research*. Washington, DC: U. S. Department of Health, Education and Welfare.
Shah, S. (2006). *The body hunters: Testing new drugs on the world's poorest patients*. New York: New Press.
Shore, N., Wong, K., Seifer, S., Grignon, J., & Gamble, V. (2008). Introduction: Advancing the ethics of community-based participatory research. *Journal of Empirical Research on Human Research Ethics*, 3, 1–4.

# NVivo

NVivo provides software tools to assist a researcher from the time of conceptualization of a project through to its completion. Although NVivo is software that is designed primarily for researchers undertaking analysis of qualitative (text and multimedia) data, its usefulness extends to researchers engaged in any kind of research. The tools provided by NVivo assist in the following:

- Tracking and management of data sources and information about these sources
- Tracking and linking ideas associated with or derived from data sources
- Searching for terms or concepts
- Indexing or coding text or multimedia information for easy retrieval
- Organizing codes to provide a conceptual framework for a study
- Querying relationships between concepts, themes, or categories
- Building and drawing visual models with links to data

Particular and unique strengths of NVivo lie in its ability to facilitate work involving complex data sources in a variety of formats, in the range of the query tools it offers, and in its ability to link quantitative with qualitative data.

The consideration of NVivo is relevant to the practical task of research design in two senses: Its tools are useful when designing or preparing for a research project, and if it is to be used also for the analysis of qualitative or mixed methods data, then consideration needs to be given to designing and planning for its use.

## Designing With NVivo

NVivo can assist in the research design process in (at least) three ways, regardless of the methodological approach to be adopted in the research: keeping a research journal, working with literature, and building conceptual models.

### *Keeping a Journal*

Keeping a record of decisions made when planning and conducting a research project, tracking events occurring during the project (foreseen and unforeseen), or even recording random thoughts about the project, will assist an investigator to prepare an accurate record of the methods adopted for the project and the rationale for those methods. Journaling serves also to simulate thinking, to prevent loss of ideas that might be worthy of follow-up, and to provide an audit trail of development in thinking toward final conclusions.

NVivo can work with text that has been recorded using Microsoft Word, or a journal can be recorded progressively within NVivo. Some researchers keep all their notes in a single document, whereas others prefer to make several documents perhaps to separate methodological from substantive issues. The critical contribution that NVivo can make here is to assist the researcher in keeping track of their ideas through coding the content of the journal. A coding system in NVivo works rather like an index but with a bonus. The text in the document is highlighted and tagged with a code (a label that the researcher devises). Codes might be theoretically based and designed a priori, or they can be created or modified (renamed, rearranged, split, combined, or content recoded) in an emergent way as the project proceeds. The bonus of using codes in NVivo is that all materials that have been coded in a particular way can be retrieved together, and if needed for clarification, any coded segment can be shown

within its original context. A research journal is a notoriously messy document, comprising random thoughts, notes of conversations, ideas from reading, and perhaps even carefully considered strategies that entered in no particular sequence. By coding their journal, researchers can find instantly any thoughts or information they have on any particular aspect of their project, regardless of how messy the original document was. This different view of what has been written not only brings order out of chaos and retrieves long-forgotten thoughts but also prompts deeper thinking and perhaps reconceptualization of that topic after visualizing all the material on one topic together.

### *Working With Literature*

NVivo's coding system can be used also to index, retrieve, and synthesize what is learned from reading across the substantive, theoretical, or methodological literature during the design phase of a project. In the same way that a journal can be coded, either notes derived from reading or the text of published articles can be coded for retrieval and reconsideration according to either an a priori or emergent system (or a combination thereof). Thus, the researcher can locate and bring together all their material from any of their references on, for example, the concept of equity, debates about the use of $R^2$, or the role of antioxidants in preventing cancer. With appropriate setting up, the author/s and year of publication for any segment of coded text can be retrieved alongside the text, which facilitates the preparation of a written review of the literature. The database created in NVivo becomes available for this and many more projects, and it serves as an ongoing database for designing, conducting, and writing up future projects.

A preliminary review of the literature can be extended into a more thorough analysis by drawing on NVivo's tools for recording and using information about sources (referred to as attributes) in comparative analyses or by examining the relationship between, say, perspectives on one topic and what is said about another. Recorded information about each reference could be used, for example, to review changes over time in perspectives on a particular concept or perhaps to compare the key concerns of North American versus European writers. The associations between codes, or between attributes and codes, could be used to review the relationship between an author's theoretical perspective and his or her understanding of the likely impact of a planned intervention.

Searching text in NVivo provides a useful supplement to coding. One could search the reports of a group of related studies, for example, for the alternative words "extraneous OR incidental OR unintended" to find anything written about the potential impact of extraneous variables on the kind of experiment being planned.

### *Building Conceptual Models*

Researchers often find it useful at the start of a project to "map" their ideas about their experimental design or about what they are expecting to find from their data gathering. Doing so can help to identify all the factors that will possibly impinge on the research process and to clarify the pathways by which different elements will impact on the process and its outcomes. As a conceptual or process model is drawn, fresh awareness of sampling or validity issues might be prompted and solutions sought. NVivo provides a modeling tool in which items and their links can be shown. A variety of shapes can be used in designing the model. Project items such as codes, cases, or attributes can be added to the model, and where coding is present, these codes provide a direct link back to the data they represent. Labels can be added to links (which might be associative, unidirectional, or bidirectional), and styles (e.g., color, fill, and font) can be used to emphasize the significance of different items or links. The items can be grouped so that they can be turned on or off in the display. The models can be archived, allowing the researcher to continue to modify their model as their understanding grows while keeping a historical record of their developing ideas.

## Designing for Analysis With NVivo

Where the intention is to use NVivo for analysis of qualitative or mixed-methods data, there are

several points to be aware of at the design stage of the project and again when preparing data to facilitate the effective (and efficient) use of NVivo's coding and query tools.

NVivo's data management system is based on the notion of the "case," with the case being best thought of as the unit of analysis for the current study. Most often, this is an individual; alternatively, the case might be a workgroup, a document, a site, or a message. The information gathered for a case might come from a single document, such as an interview or response to a survey. It might be across several documents, such as repeat interviews in a longitudinal study, or when data about the case come through multiple sources or methods; the information might come from several parts of a document, such as for a member of a focus group. All the information for a case is held together, along with attribute data relating to that case, in a node (NVivo's name for the "buckets" that hold coding)—one for each case. Additionally, sources (and nodes) can be organized in sets. Thus, for example, in a study of children with attention-deficit/hyperactivity disorder (ADHD), one might interview the children, their parents, their teachers and their doctors, observe the families in interaction, and record a number of demographic and scaled measures about each child. All the information relating to each child would be referenced in a case node for that child, allowing for within-case and cross-case analysis, but as well, one could create a set of child interviews, a set of teacher interviews, and so on. Once the content of the various data sources has also been coded for concepts and themes, the flexibility of this data management system allows for a wide range of analysis strategies.

As well as recording notes, verbatim text, or multimedia data, the astute researcher will plan for recording demographic and other categorical or scaled information when designing data collection. This information will be entered into NVivo as attributes of cases and used for comparative analyses—a valuable tool not only for comparing patterns of differences across subgroups but also for revealing subtleties within the text. Attribute data can be entered interactively within the software, or it can be entered in a spreadsheet or database as case-by-variable data and imported using a text file format.

### *Data Preparation*

Most often, data imported into NVivo are documents prepared in Word, although it is also possible to import and code video, audio and image data, and .pdf documents. There is no spell-check function in NVivo, so careful editing of the file before import is recommended (this is especially so if someone other than the interviewer has done the transcription).

Where each file to be imported represents unstructured or semistructured material for a single case, detailed preparation for importing is not necessary. Where, however, Word files each contain information for several cases, such as from a focus group or a series of internet messages, or if the individual's responses are to a series of structured questions (such as from a survey), there is much to be gained by a careful preparation of the file. Specifically, the application of styled headings which identify each case (i.e., who was speaking or whose response this was) or each question will allow for the use of an automatic coding tool to assign passages of text to individual cases, to code responses for the questions being answered, or both.

Video and audio data can be imported and coded without transcription. Alternatively, the transcription can be imported along with the multimedia file, or it can be transcribed within NVivo.

### *Designing for Integration of Qualitative and Quantitative Data and Analyses*

If the intention is to link quantitative and qualitative data for each case, it is critical to have a common identifier for both sources of data, so that it can be matched within NVivo. Quantitative data are most commonly used as a basis for subgroup comparisons of open responses, or these data can be used to define subsets of the data as the basis for specific or repeated analyses.

To go beyond simply retrieving coded text, NVivo's query function facilitates asking relational questions of the data. These include comparative and pattern analyses that are typically presented in matrix format (a qualitative cross-tabulation with both numbers and text available); questions about associations between, say, conditions and actions; questions to explore negative cases; and questions

that arise as a consequence of asking other questions (so that results from one question are fed into another). Queries can be saved so that they can be run again, with more data or with a different subset of data. Relevant text is retrieved for review and drawing inferences; patterns of coding reflected in numbers of sources, cases, or words are available in numeric form or as charts. All coding information, including results from queries, can be exported in numeric form for subsequent statistical analysis if appropriate—but always with the supporting text readily available in the NVivo database to give substance to the numbers.

*Pat Bazeley*

***See also*** Demographics; Focus Group; Interviewing; Literature Review; Mixed Methods Design; Mixed Model Design; Observations; Planning Research; Qualitative Research

### Further Readings

Bazeley, P. (2006). The contribution of computer software to integrating qualitative and quantitative data and analyses. *Research in the Schools,* 13, 64–73.
Bazeley, P. (2007). *Qualitative data analysis with NVivo.* Thousand Oaks, CA: Sage.
Richards, L. (2005). *Handling qualitative data.* Thousand Oaks, CA: Sage.

### Websites

QSR International: http://www.qsrinternational.com
Research Support Pty. Limited: http://www.researchsupport.com.au

## Observational Research

The observation of human and animal behavior has been referred to as the sine qua non of science, and indeed, any research concerning behavior ultimately is based on observation. A more specific term, *naturalistic observation,* traditionally has referred to a set of research methods wherein the emphasis is on capturing the dynamic or temporal nature of behavior in the environment where it naturally occurs, rather than in a laboratory where it is experimentally induced or manipulated. What is unique about the more general notion of observational research, however, and what has made it so valuable to science is the fact that the process of direct systematic observation (that is, the *what, when, where,* and *how* of observation) can be controlled to varying degrees, as necessary, while still permitting behavior to occur naturally and over time. Indeed, the control of what Roger Barker referred to as "the stream of behavior," in his 1962 book by that title, may range from a simple specification of certain aspects of the context for comparative purposes (e.g., diurnal vs. nocturnal behaviors) to a full experimental design involving the random assignment of participants to strictly specified conditions.

Even the most casual observations have been included among these research methods, but they typically involve, at a minimum, a systematic process of specifying, selecting, and sampling behaviors for observation. The behaviors considered might be maximally inclusive, such as in the case of the *ethogram,* which attempts to provide a comprehensive description of all of the characteristic behavior patterns of a species, or they might be restricted to a much smaller set of behaviors, such as the social behaviors of jackdaws, as studied by the Nobel Prize–winning ethologist Konrad Lorenz, or the facial expressions of emotion in humans, as studied by the psychologist Paul Ekman. Thus, the versatile set of measurement methods referred to as *observational research* emphasizes temporally dynamic behaviors as they naturally occur, although the conditions of observation and the breadth of behaviors observed will vary with the research question(s) at hand.

Because of the nature of observational research, it is often better suited to hypothesis generation than to hypothesis testing. When hypothesis testing does occur, it is limited to the study of the relationship(s) between/among behaviors, rather than to the causal links between them, as is the focus of experimental methods with single or limited behavioral observations and fully randomized designs. This entry discusses several aspects of observational research: its origins, the approaches, special considerations, and the future of observational research.

### Origins

Historically, observational research has its roots in the naturalistic observational methods of Charles Darwin and other naturalists studying nonhuman

animals. The work of these 19th-century scientists spawned the field of ethology, which is defined as the study of the behavior of animals in their natural habitats. Observational methods are the primary research tools of the ethologist. In the study of human behaviors, a comparable approach is that of ethnography, which combines several research techniques (observations, interviews, and archival and/or physical trace measures) in a long-term investigation of a group or culture. This technique also involves immersion and even participation in the group being studied in a method commonly referred to as participant observation.

The use of observational research methods of various kinds can be found in all of the social sciences—including, but not limited to, anthropology, sociology, psychology, communication, political science, and economics—and in fields that range from business to biology, and from education to entomology. These methods have been applied in innumerable settings, from church services to prisons to psychiatric wards to college classrooms, to name a few.

## Distinctions Among Methods

Whether studying humans or other animals, one of the important distinctions among observational research methods is whether the observer's presence is overt or *obtrusive* to the participants or covert or *unobtrusive*. In the former case, researchers must be wary of the problem of *reactivity of measurement;* that is, of measurement procedures where the act of measuring may, in all likelihood, change the behavior being measured. Reactivity can operate in a number of ways. For example, the physical space occupied by an observer under a particular tree or in the corner of a room may militate against the occurrence of the behaviors that would naturally occur in that particular location. More likely, at least in the case of the study of human behavior, participants may attempt to control their behaviors in order to project a certain image. One notable example in this regard has been termed *evaluation apprehension*. Specifically, human participants who know that they are being observed might feel apprehensive about being judged or evaluated and might attempt to behave in ways that they believe put them in the most positive light, as opposed to behaving as they would naturally in the absence of an observer. For example, anthropological linguists have observed the hypercorrection of speech pronunciation "errors" in lower- and working-class women when reading a list of words to an experimenter compared to when speaking casually. Presumably, compared to upper-middle-class speakers, they felt a greater need to "speak properly" when it was obvious that their pronunciation was the focus of attention. Although various techniques exist for limiting the effects of evaluation apprehension, obtrusive observational techniques can never fully guarantee the nonreactivity of their measurements.

In the case of unobtrusive observation, participants in the research are not made aware that they are being observed (at least not at the time of observation). This can effectively eliminate the problem of measurement reactivity, but it presents another issue to consider when the research participants are humans; namely, the ethics of making such observations. In practice, ethical considerations have resulted in limits to the kinds of behaviors that can be observed unobtrusively, as well as to the techniques (for example, the use of recording devices) that can be employed. If the behavior occurs in a public place where the person being observed cannot reasonably expect complete privacy, the observations may be considered acceptable. Another guideline involves the notion of *minimal risk*. Generally speaking, procedures that involve no greater risk to participants than they might encounter in everyday life are considered acceptable. Before making unobtrusive observations, researchers should take steps to solicit the opinions of colleagues and others who might be familiar with issues of privacy, confidentiality, and minimal risk in the kinds of situations involved in the research. Research conducted at institutions that receive federal funding will have an institutional review board composed of researchers and community members who review research protocols involving human participants and who will assist researchers in determining appropriate ethical procedures in these and other circumstances.

## Special Considerations

Observational research approaches generally include many more observations or data points than typical experimental approaches, but they, too, are

reductionistic in nature; that is, although relatively more behaviors are observed and assessed, not all behaviors that occur during data collection may be studied. This fact raises some special considerations.

### *How Will the Behaviors Being Studied Be Segmented?*

Aristotle claimed that "natural" categories are those that "carve at the joint." Some behaviors do seem to segment relatively easily via their observable features, such as speaking turns in conversation, or the beginning and end of an eye blink. For many other behaviors, beginnings and endings may not be so clear. Moreover, research has shown that observers asked to segment behaviors into the smallest units they found to be natural and meaningful formed different impressions than observers asked to segment behaviors into the largest units they found natural and meaningful, despite observing the same videotaped series of behaviors. The small-unit observers also were more confident of their impressions. Consumers of observational research findings should keep in mind that different strategies for segmenting behavior may result in different kinds of observations and inferences.

### *How Will Behavior Be Classified or Coded?*

The central component of all observational systems is sometimes called a behavior code, which is a detailed description of the behaviors and/or events to be observed and recorded. Often, this code is referred to as a *taxonomy* of behavior. The best taxonomies consist of a set of categories with the features of being *mutually exclusive* (that is, every instance of an observed behavior fits into one and only one category of the taxonomy) and *exhaustive* (that is, every instance of an observed behavior fits into one of the available categories of the taxonomy).

### *Are the Classifications of Observed Behaviors Reliable Ones?*

The coding of behaviors according to the categories of a taxonomy have, as a necessary condition, that the coding judgments are reliable ones. In the case of *intrarater reliability,* this means that an observer should make the same judgments regarding behavior codes if the behaviors are observed and classified again at another time. In the case of *interrater reliability,* two (or more) judges independently viewing the behaviors should make the same classifications or judgments. Although in practice, reliability estimates seldom involve perfect agreement between judgments made at different times or by different coders, there are standards of disagreement accepted by researchers based upon the computations of certain descriptive and inferential statistics. The appropriate statistic(s) to use to make a determination of reliability depends upon the nature of the codes/variables being used. Correlations often are computed for continuous variables or codes (that is, for classifications that vary along some continuum; for example, degrees of displayed aggression), and Cohen's kappa coefficients often are computed for discrete or categorical variables or codes; for example, types of hand gestures.

### *What Behaviors Will Be Sampled?*

The key to sampling is that there is a sufficient amount and appropriate kind of sampling performed such that one represents the desired population of behaviors (and contexts and types of participants) to which one would want to generalize. Various sampling procedures exist, as do statistics to help one ascertain the number of observations necessary to test the reliability of the measurement scheme employed and/or test hypotheses about the observations (for example, power analyses and tests of effect size).

## Problems Associated With Observational Research

Despite all of the advantages inherent in making observations of ongoing behavior, a number of problems are typical of this type of research. Prominent among them is the fact that the development and implementation of reliable codes can be time-consuming and expensive, often requiring huge data sets to achieve representative samples and the use of recording equipment to facilitate reliable measurement. Special methods may be needed to prevent, or at least test for, what has been called *observer drift.* This term refers to the fact that, with prolonged observations, observers may be more likely to forget coding details,

become fatigued, experience decreased motivation and attention, and/or learn confounding habits. Finally, observational methods cannot be applied to hypotheses concerning phenomena not susceptible to direct observation, such as cognitive or affective variables. Indeed, care must be taken by researchers to be sure that actual observations (e.g., he smiled or the corners of his mouth were upturned or the zygomaticus major muscle was contracted) and not inferences (e.g., he was happy) are recorded as data.

### Future Outlook

With the increasing availability and sophistication of computer technology, researchers employing observational research methods have been able to search for more complicated patterns of behavior, not just within an individual's behavior over time, but among interactants in dyads and groups as well. Whether the topic is family interaction patterns, courtship behaviors in Drosophila, or patterns of nonverbal behavior in doctor-patient interactions, a collection of multivariate statistical tools, including factor analyses, time-series analyses, and t-pattern analyses, has become available to the researcher to assist him or her in detecting the hidden yet powerful patterns of behavior that are available for observation.

*Carol Toris*

***See also*** Cause and Effect; Cohen's Kappa; Correlation; Effect Size, Measures of; Experimental Design; Hypothesis; Laboratory Experiments; Multivariate Analysis of Variance (MANOVA); Naturalistic Observation; Power Analysis; Reactive Arrangements; Reliability; Sample Size Planning; Unit of Analysis

#### Further Readings

Barker, R. G. (Ed.). (1963). *The stream of behavior: Explorations of its structure and content*. New York: Appleton-Century-Crofts.

Campbell, D. T., & Stanley, J. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Ekman, P. (1982). Methods for measuring facial action. In K. R. Scherer & P. Ekman (Eds.), *Handbook of methods in nonverbal behavior research*. Cambridge, UK: Cambridge University Press.

Hawkins, R. P. (1982). Developing a behavior code. In D. P. Hartmann (Ed.), *Using observers to study behavior.* San Francisco: Jossey-Bass.

Jones, R. (1985). *Research methods in the social and behavioral sciences*. Sunderland, MA: Sinauer Associates, Inc.

Longabaugh, R. (1980). The systematic observation of behavior in naturalistic settings. In H. Triandis (Ed.), *The handbook of cross-cultural psychology: II, Methodology.* Boston: Allyn & Bacon.

Magnusson, M. S. (2005). Understanding social interaction: Discovering hidden structure with model and algorithms. In L. Anolli, S. Duncan, Jr., M. S. Magnusson, & G. Riva (Eds.), *The hidden structure of interaction.* Amsterdam: IOS Press.

Suen, H. K., & Ary, D. (1989). *Analyzing quantitative behavioral observation data.* Mahwah, NJ: Lawrence Erlbaum.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals. *American Psychologist, 54,* 594–604.

## Observations

Observations refer to watching and recording the occurrence of specific behaviors during an episode of interest. The observational method can be employed in the laboratory as well as a wide variety of other settings to obtain a detailed picture of how behavior unfolds. This entry discusses types of observational design, methods for collecting observations, and potential pitfalls that may be encountered.

### Types of Observational Designs

There are two types of observational design: naturalistic and laboratory observations. Naturalistic observations entail watching and recording behaviors in everyday environments such as animal colonies, playgrounds, classrooms, and retail settings. The main advantage of naturalistic observation is that it affords researchers the opportunity to study the behavior of animals and people in their natural settings. Disadvantages associated with naturalistic observations are lack of control over the setting; thus, confounding factors may come into play. Also, the behavior of interest may be extremely

infrequent and unlikely to be captured during observational sessions.

Laboratory observations involve watching and recording behaviors in a laboratory setting. The advantage of laboratory observations is that researchers can structure them to elicit certain behaviors by asking participants to discuss a particular topic or complete a specific task. The major disadvantage is that participants may behave unnaturally because of the contrived nature of the laboratory.

## Collecting Observations

### *Specifying the Behavior of Interest*

The first step in collecting observations is to specify the behavior(s) of interest. This often consists of formulating an operational definition, or precisely describing what constitutes an occurrence of each type of behavior. For instance, physical aggression may be operationally defined as hitting, kicking, or biting another person. Thus, when any of these behaviors occur, the researcher would record an instance of physical aggression. Researchers often create coding manuals, which include operational definitions and examples of the behaviors of interest, to use as a reference guide when observing complex behaviors.

### *Recording Observations*

Researchers use different methods for recording observations dependent on the behavior of interest and the setting. One dimension that may differ is whether observations are made live or while watching a recording of the episode. Researchers often choose to record simple behaviors live and in real time as they occur. In situations characterized by complexity, researchers often elect to make a DVD/video recording of the episode. This gives researchers more flexibility in recording a variety of behaviors, most notably the ability to progress at their own speed, or to watch behaviors again if needed.

Similarly, researchers may choose to record behavior by making hand tallies or using computers. If the behavior interest is simple, researchers may choose to record each time a behavior occurs. Many researchers, however, choose to use a computer to facilitate observational research. Several computer programs are available for recording observations. Computer entry allows for exact timing of behavior so that researchers can determine time lags between particular instances of behavior.

Another choice in recording behavior is whether to use time or event sampling. Time sampling involves dividing the observational time session into short time periods and recording any occurrences of the behavior of interest. For instance, researchers studying classroom participation might divide a 1-hour class into twelve 5-minute intervals. They could then record if students participated in each interval and then determine the percentage of intervals that included student participation. A second option is to use event sampling, recording each behavior of interest as it occurs. Researchers using the event sampling technique in the classroom participation example would record each time a student participated in class and would ultimately calculate the frequency of student participation across the class period.

### *Ensuring Interrater Reliability*

Interrater reliability refers to agreement among observers and is necessary for scientifically sound observations. Several steps can be taken to ensure high interrater reliability. First, new observers often receive detailed training before they begin to code observational data. In many cases, novice observers also practice with DVD/video recordings to gain experience with the observational task. Finally, two or more observers often code a percentage of the episodes to ensure that agreement remains high.

## Potential Pitfalls

There are two potential dangers in collecting observations: observer influence and observer bias. Observer influence refers to changes in behavior in response to the presence of an observer. Individuals may be cognizant of the observer and alter their behavior, often in a more positive direction. The term *Hawthorne effect* is often used to refer to these increases in positive (e.g., prosocial, productive) behavior in response to an observer. Steps can be taken in order to reduce observer influence,

including utilization of adaptation periods during which observers immerse themselves in the environment prior to data collection so that the subjects of their observation become accustomed to their presence.

Another potential danger is observer bias, in which observers' knowledge of the study hypotheses influences their recording of behavior. Observers may notice and note more behavior that is congruent with the study hypotheses than actually occurs. At the same time, they may not notice and note behavior that is incongruent with the study hypotheses. One means of lessening observer bias is to limit the information given to observers regarding the study hypotheses.

*Lisa H. Rosen and Marion K. Underwood*

***See also*** Hawthorne Effect; Interrater Reliability; Naturalistic Observation; Observational Research

**Further Readings**

Dallos, R. (2006). Observational methods. In G. Breakwell, S. Hammond, C. Fife-Schaw, & J. Smith (Eds.), *Research methods in psychology* (pp. 124–145). Thousand Oaks, CA: Sage.

Margolin, G., Oliver, P. H., Gordis, E. B., O'Hearn, H. G., Medina, A. M., Ghosh, C. M., & Morland, L. (1998). The nuts and bolts of behavioral observation of marital and family interaction. *Clinical Child and Family Psychology Review, 1,* 195–213.

Pope, C., & Mays, N. (2006). Observational methods. In C. Pope & N. Mays (Eds.), *Qualitative research in health care* (pp. 32–42). Malden, MA: Blackwell.

# Occam's Razor

Occam's Razor (also spelled Ockham) is known as the *principle of parsimony* or the *economy of hypotheses*. It is a philosophical principle dictating that, all things being equal, simplicity is preferred over complexity. Traditionally, the Razor has been used as a philosophical heuristic for choosing between competing theories, but the principle is also useful for defining methods for empirical inquiry, selecting scientific hypotheses, and refining statistical models. According to Occam's Razor, a tool with fewer working parts ought to be selected over one with many, provided they are equally functional. Likewise, a straightforward explanation ought to be believed over one that requires many separate contingencies.

For instance, there are a number of possible reasons why a light bulb does not turn on when a switch is flipped: Aliens could have abducted the light bulb, the power could be out, or the filament within the bulb has burned out. The explanation requiring aliens is exceedingly complex, as it necessitates the existence of an unknown life form, a planet from which they have come, a motive for taking light bulbs, and so on. A power outage is not as complicated, but still requires an intricate chain of events, such as a storm, accident, or engineering problem. The simplest of these theories is that the light bulb has simply burned out. All theories provide explanations, but vary in complexity. Until proof corroborating one account surfaces, Occam's Razor requires that the simplest explanation be preferred above the others. Thus, the logical—and most likely correct—hypothesis is that the light bulb has burned out.

This entry begins with a brief history of Occam's Razor. It then discusses the implications for research. The entry concludes with some caveats related to the use of Occam's Razor.

## History

Occam's Razor is named for the 14th-century English theologian, philosopher, and friar William of Occam. William, who was presumably from the city of Occam, famously suggested that "entities should not be multiplied beyond necessity." To do so, he explained, implied vanity and needlessly increased the chances of error. This principle had been formalized since the time of Aristotle, but Occam's unabashed and consistent use of the Razor helped Occam become one of the foremost critics of Thomas Aquinas.

## Implications for Scientific Research

The reasons for emphasizing simplicity when conceptualizing and conducting research may seem obvious. Simple designs reduce the chance of experimenter error, increase the clarity of the results, obviate needlessly complex statistical analyses, conserve valuable resources, and curtail

potential confounds. As such, the Razor can be a helpful guide when attempting to produce an optimal research design. Although often implemented intuitively, it may be helpful to review and refine proposed research methods with the Razor in mind.

Just as any number of tools can be used to accomplish a particular job, there are many potential methodological designs for each research question. Occam's Razor suggests that a tool with fewer working parts is preferable to one that is needlessly complicated. A correlation design that necessitates only examining government records may be more appropriate than an experimental design that necessitates recruiting, assigning, manipulating, and debriefing participants. If both designs would yield the same conclusion, Occam's Razor dictates that the simpler correlation design should be used.

Although the Razor is implicitly useful in selecting a design, its overt utility can be seen when streamlining research. Before beginning any research endeavor, it is wise to determine the value of each component. Occam's Razor reasonably warns against the implementation of any component that needlessly complicates the enterprise. Although the use of new technology, participant deception, or novel methodology may be considered "sexy" or "fun" research, the inclusion of such factors may sacrifice the validity or reliability of the investigation.

The principle of Occam's Razor is also apparent in some statistical analyses. A regression analysis is a statistical tool that allows a researcher to assess the contribution of multiple independent variables on a single dependent variable. After determining the relative contribution of each independent variable, the researcher must adjust his or her model to include only those variables that add enough descriptive information to warrant their inclusion. A regression model with two independent variables should not include a third unless it accounts for additional unique variance. A model with two independent variables, for instance, that account for 80% of the variance of the dependent variable is preferable to a model with three variables that account for the same 80%. Its inclusion precisely mimics the needless multiplication of entities prohibited by Occam. A model with three independent variables is considerably more complex than a model with two independent variables. Because the third variable does not contribute information to the model, Occam's Razor can be used to cut it away.

## Caveats

In practice, a strict adherence to Occam's Razor is usually impossible or ill-advised, as it is rare to find any two models or theories that are equivalent in all ways except complexity. Often, when some portion of a method, hypothesis, or theory is cut away, some explanatory or logical value must be sacrificed. In the previously mentioned case of the regression analysis, the addition of a third independent variable may contribute a small amount of descriptive value, perhaps .01%. It is unlikely that a variable accounting for .01% of the variance will significantly affect a model and deem inclusion. If it accounted for 5%, it would be more tempting to accept a more complicated model by allowing it to remain; and if it accounted for 10%, it would certainly pass the Razor without being sliced away.

Furthermore, it should not be assumed that one extremely complicated hypothesis or variable is preferred over two simple ones. For instance, a model using one variable, *weather,* may be just as predictive as a model with two variables, *temperature* and *humidity.* Although it appears that the first model is the simpler—and therefore preferable by Occam's Razor—the variable, *weather,* is likely to contain many discrete subvariables such as cloud cover, barometric pressure, and wind, as well as temperature and humidity. Although simply stated, *weather* is not a simple concept. It is important to remember that Occam's Razor advocates simplicity, which is not to be confused with brevity.

## Conclusion

At its best, the Razor is used to trim unnecessary complexities from any theory or design, leaving only that which can be directly used to answer the research question. Correct use of the Razor at each step of the research process will reduce any final product to its most efficient, elegant, and consumable form, maximizing accuracy and reducing error.

*Thomas C. Motl*

*See also* Bivariate Regression; Dependent Variable; Hypothesis; Independent Variable

**Further Readings**

Fearn, N. (2001). Ockham's Razor. In N. Fearn (Ed.), *Zeno and the tortoise: How to think like a philosopher* (pp. 56–60). New York: Grove.

Kaye, S. M., & Martin, R. M. (2000). *On Ockham*. New York: Wadsworth.

# Odds

Let the probability of the occurrence of an event be denoted by $p$, thereby implying that $1 - p$ represents the probability of the nonoccurrence of the event. Succinctly, the *odds* of the occurrence of the event are defined as the ratio of the probability of the occurrence to the probability of the nonoccurrence, or $\frac{p}{1-p}$.

Odds are often encountered in gambling situations. A pertinent example can be extracted from a football game. For instance, if a Vegas odds maker assigns an 80% probability of a football team from western New York winning a match against another team from eastern Massachusetts, the suggested odds of the football team from western New York winning are $\frac{0.8}{1-0.8} = 4.0$.

If the odds can be represented in terms of a fraction $\frac{a}{b}$, where $a$ and $b$ are both integers, the odds are often referred to as "$a$-to-$b$." Thus, the odds in favor of the team from western New York winning the match would be quoted as "four-to-one."

In the context of gambling, a "fair game" is one in which both the house and the gambler have an expected gain of zero. Indeed, if the odds maker sets the odds of the team from western New York winning at four-to-one, and the perceived odds among gamblers are consistent with this assignment, then the average gain for both the house and the gamblers should be near zero. In this setting, if a gambler has the discretionary income to bet \$1 in favor of the eastern Massachusetts football team, and the team from eastern Massachusetts wins the match, the gambler would receive $4 \times \$1 = \$4$ in earnings plus all monies with which the bet was made (\$1), thus accumulating a total of \$5. Conversely, should the gambler choose to bet \$1 in favor of the western New York football team, and the team from western New York prevails, the gambler would receive $\frac{1}{4} \times \$1 = \$0.25$ in earnings plus all monies with which the bet was placed (\$1), thus accumulating a total of \$1.25. Neither the gambler nor the house experiences a positive expected gain because the odds of the team from western New York defeating the team from eastern Massachusetts directly offset the difference in payouts.

In biostatistics, odds are often used to characterize the likelihood of a disease or condition. For example, if 100 people within a small village of 1,100 people contract a certain disease, then the odds of a randomly selected resident contracting the disease are $\frac{100}{1,000} = 0.1$. An interpretation of this measure would be that the probability of becoming infected with the disease is 0.1 times the probability of not developing the disease.

Care must be taken to distinguish between odds and probability, especially when the measures are both between 0 and 1. Consider, for instance, tossing an unbiased, six-sided die. The probability of rolling a 1 is $\frac{1}{6}$; however, the odds of rolling a 1 are $\frac{1}{5}$, or "one-to-five."

For any event that is not impossible, the odds are always greater than the probability. The previous examples illustrate this fact. The probability of a western New York football team victory is set at 0.8, whereas the odds are 4.0. The probability of a villager contracting the disease is 0.091, whereas the odds are 0.1. The probability of rolling a 1 in a die toss is 0.167, whereas the odds are 0.2.

Probability is bounded between zero and one. In other words, the probability of the occurrence of an event, $p$, is such that $0 \leq p \leq 1$. Odds are bounded between zero and infinity, or

$$0 \leq \frac{p}{1-p} < \infty,$$

when $0 \leq p < 1$.

When the probability of the occurrence of an event is small, the odds and the probability are very similar. This is evident in the disease example. The probability of the event (becoming diseased) is small, and so the probability, 0.091, and the odds, 0.1, are quite close.

If $p_1$ represents the probability of an event occurring and $p_2$ represents the probability of

a second event occurring, the ratio of the odds of the two events, known as the *odds ratio*, is

$$\frac{p_1}{1-p_1} / \frac{p_2}{1-p_2}.$$

Often, in comparing the odds for two events, the individual odds are not reported, but rather the odds ratio is used. Another common measure based on the odds is the log of the odds,

$$\log\left(\frac{p}{1-p}\right),$$

which is commonly referred to as the logit and plays an integral part in *logistic regression*. The logit is unbounded and may assume any value:

$$-\infty < \log\left(\frac{p}{1-p}\right) < \infty.$$

It thereby provides a more convenient measure to model than the odds or the probability, both of which are bounded.

*Joseph E. Cavanaugh and Eric D. Foster*

***See also*** Logistic Regression; Odds Ratio

**Further Readings**

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). Hoboken, NJ: Wiley.
Rosner, B. (2006). *Fundamentals of biostatistics*. Belmont, CA: Thomson Brooks/Cole.

# Odds Ratio

The odds ratio (OR) is a measure of association that is used to describe the relationship between two or more categorical (usually dichotomous) variables (e.g., in a contingency table) or between continuous variables and a categorical outcome variable (e.g., in logistic regression). The OR describes how much more likely an outcome is to occur in one group as compared to another group. ORs are particularly important in research settings that have dichotomous outcome variables (e.g., in medical research).

As the name implies, the OR is the ratio of two "odds," which are, in turn, ratios of the chance or probability of two (or more) possible outcomes. Suppose one were throwing a single die and wanted to calculate the odds of getting a 1, 2, 3, or 4. Because there is a 4 in 6 chance of throwing a 1 through 4 on a single die, the odds are 4/6 (the probability of getting a 1 through 4) divided by 2/6 (2/6 is the probability of not getting a 1 through 4, but rather getting a 5 or 6) or 4/2 = "2 to 1" = 2.

## An Example: Odds Ratios for Two Dichotomous Outcomes

David P. Strachan, Barbara K. Butland, and H. Ross Anderson report on the occurrence of hay fever for 11-year-old children with and without eczema and present the results in a contingency table (Table 1).

First, the probability of hay fever for those with eczema (the top row) is calculated. This probability is 141/561 = .251. Thus, the odds of hay fever for those with eczema are

$$\frac{\frac{141}{561}}{\frac{420}{561}} = \frac{141}{420} = .34,$$

about 1 to 3. Analogously, the probability of hay fever for those without eczema is $\frac{928}{14,453} = .064$, and the odds are $\frac{928}{13,525}$. The OR in this example is defined as the odds of hay fever for eczema patients divided by the odds of hay fever for noneczema patients. The OR, therefore, is

$$\frac{\frac{141}{420}}{\frac{928}{13,525}} = \frac{141 \times 13,525}{420 \times 928} = 4.89. \qquad (1)$$

From the example, one can infer that the odds of hay fever for eczema sufferers are 4.89 times the odds for noneczema patients. Thus, having eczema almost quintuples the odds of getting hay fever.

**Table 1** Contingency Table of Observed Numbers of Cases of Eczema and Hay Fever

| | *Hay Fever* | | |
|---|---|---|---|
| *Eczema* | *Yes* | *No* | *Total* |
| Yes | 141 | 420 | 561 |
| No | 928 | 13,525 | 14,453 |
| Total | 1,069 | 13,945 | 15,522 |

The ratio of the two odds can, as shown in Equation (1), also be computed as a ratio of the products of the diagonally opposite cells and is also referred to as the *cross-product ratio*.

The OR is bounded by zero and positive infinity. Values above and below 1 indicate that the occurrence of an event is more likely for one or the other group, respectively. An OR of exactly 1 means that the two odds are exactly equal, implying complete independence between the variables.

## Does the Odds Ratio Imply a Significant Relationship?

To determine whether or not this OR is significantly different from 1.0 (implying the observed relationship or effect is most likely not due to chance), one can perform a null hypothesis significance test. Usually, the OR is first transformed into the *log odds ratio* [log(OR)] by taking its natural logarithm. Then, this value is divided by its standard error and the result compared to a test value. In this example, the OR of 4.89 is transformed to the log(OR) of 1.59. A log(OR) of 0 implies independence, whereas values further away from 0 signify relationships in which the probability or odds are different for the two groups. Note that log(OR) is symmetrical and is bounded by negative and positive infinity.

The standard error (SE) of the log(OR) is defined as

$$SE_{\log(\mathrm{OR})} = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}\ , \qquad (2)$$

where $n_{11}$ to $n_{22}$ are sample sizes of the four cells in the contingency table. Using this formula, the $\mathrm{SE}_{\log(\mathrm{OR})}$ in the example is

$$SE_{\log(\mathrm{OR})} = \sqrt{\frac{1}{141} + \frac{1}{420} + \frac{1}{928} + \frac{1}{13,525}} = 0.103.$$

Thus, $\frac{\log(\mathrm{OR})}{SE_{\log(\mathrm{OR})}} = \frac{1.59}{.103} = 15.44$.

In large samples, the sampling distribution of log(OR) can be assumed to be normal and thus can be compared to the critical values of the normal distribution (that is, with $\alpha = .05$, the values +/− 1.96). Because the obtained value far exceeds this critical value, the OR can be said to be significantly different from 1 at the .05 level.

Alternatively, one can use the confidence interval approach. The 95% confidence interval for this example's log(OR) is 1.59 +/− (0.103)(1.96) and thus ranges from 1.386 to 1.790. The result of the previous significance test is reflected here by the fact that the interval does not contain 0. Alternatively, the 95% confidence interval around the OR [rather than the log(OR)] can be found. To do so, one takes the antilog of the confidence limits of the log(OR) interval, yielding a confidence interval around the OR estimate of 4.89 that ranges from 4.00 to 5.99. Because the OR was shown earlier to be statistically significant, the interval will not contain 1.0. Note that the confidence interval around the log(OR) is symmetrical, whereas that of the OR is not.

## Odds Ratios for Continuous Variables and a Categorical Outcome

When a categorical outcome is to be predicted from several variables, either categorical or continuous, it is common to use logistic regression. The results of a logistic regression are often reported in the form of an OR or log(OR). The interpretation of these coefficients changes slightly in the presence of continuous variables. For example, Dean G. Kilpatrick and colleagues examined risk factors for substance abuse in adolescents. Among many other findings, they report that age was a significant predictor of whether or not a diagnosis of substance abuse was given or not. The logistic parameter coefficient predicting substance abuse from age is 0.67. The OR can be derived from that by simply raising $e$ to the power of this number. The OR is therefore $e^{0.67} = 1.95$. Thus, each additional year of age yields 1.95 times the odds of being diagnosed with a substance abuse problem. ORs do not increase linearly, but exponentially, meaning that an additional 5 years of age has $e^{(0.67)*5} = e^{3.35} = 28.5$ times the odds.

*Sanford L. Braver, Felix Thoemmes, and Stephanie E. Moser*

***See also*** Categorical Variable; Confidence Intervals; Logistic Regression; Normal Distribution; Odds

**Further Readings**

Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed.). New York: Wiley.

Bland, J. M., & Altman, D. G. (2000). The odds ratio. *British Medical Journal, 320,* 1468.

Kilpatrick, D., Ruggiero, K., Acierno, R., Saunders, B., Resnick, H., & Best, C. (2003). Violence and risk of PTSD, major depression, substance abuse/dependence, and comorbidity: Results from the national survey of adolescents. *Journal of Consulting and Clinical Psychology, 71,* 692–700.

Strachan, D. P., Butland, B. K., & Anderson, H. R. (1996). Incidence and prognosis of asthma and wheezing illness from early childhood to age 33 in a national British cohort. *British Medical Journal, 312,* 1195–1199.

# OGIVE

Ogives are also known as cumulative frequency polygons because they are drawn on the basis of cumulative frequencies. They graphically show the total in a distribution at any given time. Ogives may also be used to determine where a particular observation stands in relation to all the other observations in the analyzed sample or population. In other words, they are useful in calculating percentiles and percentile ranks, particularly the median, the first and third quartiles, and the interquartile range (IQR). They may also be used for comparing data from two or more different samples or populations. This entry focuses on the process of constructing an ogive for both ungrouped and grouped data and on the usage of ogives to calculate percentiles and percentile ranks.

## Ogives With Ungrouped Data

Ungrouped data refer to raw data that have not been classified into categories. For example, the scores obtained by students on a final exam are considered in Table 1.

A frequency table is needed to construct an ogive from the raw data. The frequency table may include the cumulative frequencies, the relative cumulative frequencies, and/or the percentage cumulative frequencies for the analyzed data. In the case of the scores obtained by students on a final exam, the cumulative frequencies show the number of students who scored up to a particular number of points. The relative/percentage cumulative frequencies show the proportion/percentage of students who scored up to a particular number of points (Table 2).

**Table 1** Ascending Array of the Scores Obtained by Students on a Final Exam

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 61 | 61 | 65 | 65 | 75 | 78 | 78 | 84 | 84 | 85 |
| 85 | 89 | 89 | 89 | 91 | 95 | 95 | 95 | 98 | 98 |

**Table 2** Frequency Table of the Scores Obtained by Students on a Final Exam

| *Score (1)* | *Frequency (2)* | *Cumulative Frequency (3)* | *Relative Cumulative Frequency (4)* | *Percentage Cumulative Frequency (5)* |
|---|---|---|---|---|
| 61 | 2 | 2 | 0.10 | 10% |
| 65 | 2 | 4 | 0.20 | 20% |
| 75 | 1 | 5 | 0.25 | 25% |
| 78 | 2 | 7 | 0.35 | 35% |
| 84 | 2 | 9 | 0.45 | 45% |
| 85 | 2 | 11 | 0.55 | 55% |
| 89 | 3 | 14 | 0.70 | 70% |
| 91 | 1 | 15 | 0.75 | 75% |
| 95 | 3 | 18 | 0.90 | 90% |
| 98 | 2 | 20 | 1.00 | 100% |
| Total | 20 | | | |

The cumulative frequency distribution is obtained by adding the frequency corresponding to each score (column 2 in Table 2) to the sum of the frequencies of all smaller scores (column 3 in Table 2). For example, the cumulative frequency for the score of 61 is $0 + 2 = 2$, because this is the lowest score obtained by students. The cumulative frequency for the next score (65) is $2 + 2 = 4$. The next cumulative frequency (corresponding to a score of 75) is $4 + 1 = 5$, and so on. This means that, for example, 5 students scored 75 points or less on the final exam, whereas 18 students scored 95 points or less.

The relative cumulative frequencies are obtained by dividing each cumulative frequency by the total number of observations (20 in this example). The percentage cumulative frequencies are obtained by multiplying the relative cumulative frequency by 100. In relative cumulative frequency terms, it can be concluded from Table 2 that, for example,

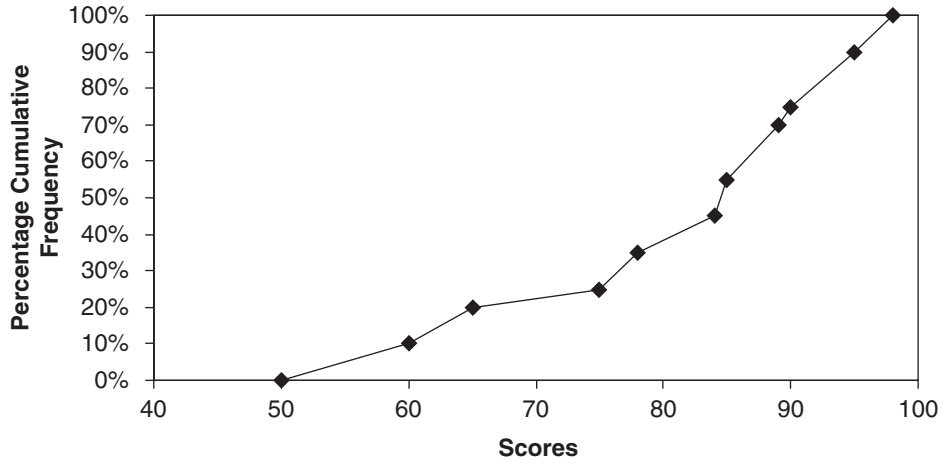



**Figure 1** Ogive of the Scores Obtained by Students on a Final Exam, Using Ungrouped Data

a proportion of 0.20 of students scored 65 points or less on the final exam. Expressed in percentage, 20% of the students scored 65 points or less.

The cumulative frequency of the highest score always equals the total number of observations (the sum of all frequencies, 20 in this example). The corresponding relative cumulative frequency is always 1.00; the corresponding percentage cumulative frequency is always 100%.

An ogive for the data in Table 2 is drawn using the scores obtained by students on the exam on the *x*-axis and either the cumulative, the relative, or the percentage cumulative frequencies on the *y*-axis (Figure 1).

The ogive is read in the same manner as the cumulative frequencies in Table 2. Reading from left to right, the ogive always remains level or increases and can never drop down toward the *x*-axis. This is because cumulative frequencies are obtained by successive additions, so that the cumulative frequency for a score can be at most equal to, but never less than, the cumulative frequency for the preceding score. A steeper slope on a certain segment of the ogive indicates a greater increase than a more gradual slope. By the time the highest score obtained on the exam is reached, the ogive reaches 100% on the *y*-axis.

## Ogives With Grouped Data

Grouped data refer to data that have been classified into categories. It is important to know how to use grouped data in order to draw ogives, because larger samples are often presented for simplicity in grouped form. The scores obtained by students on a final exam may be grouped as in Table 3.

An approach similar to the one described in the case of ungrouped data is used to obtain the cumulative frequencies. The cumulative frequency distribution is obtained by adding the frequency corresponding to each group to the sum of the frequencies of all lower range groups. For example, the cumulative frequency for the group 61–70 is $0 + 4 = 4$, because this is the group with the lowest scores obtained by students. The cumulative frequency for the next group (71–80) is $4 + 3 = 7$. The next cumulative frequency (corresponding to the group 81–90) is $7 + 7 = 14$, and so on. This means that, for example, 14 students scored 90 points or less on the final exam.

**Table 3** Grouped Frequencies Table for the Scores Obtained by Students on a Final Exam

| *Score Groups* | *Frequency* | *Cumulative Frequency* | *Relative Cumulative Frequency* | *Percentage Cumulative Frequency* |
|---|---|---|---|---|
| ≤ 60 | 0 | 0 | 0.00 | 0% |
| 61–70 | 4 | 4 | 0.20 | 20% |
| 71–80 | 3 | 7 | 0.35 | 35% |
| 81–90 | 7 | 14 | 0.70 | 70% |
| 91–100 | 6 | 20 | 1.00 | 100% |
| Total | 20 | | | |

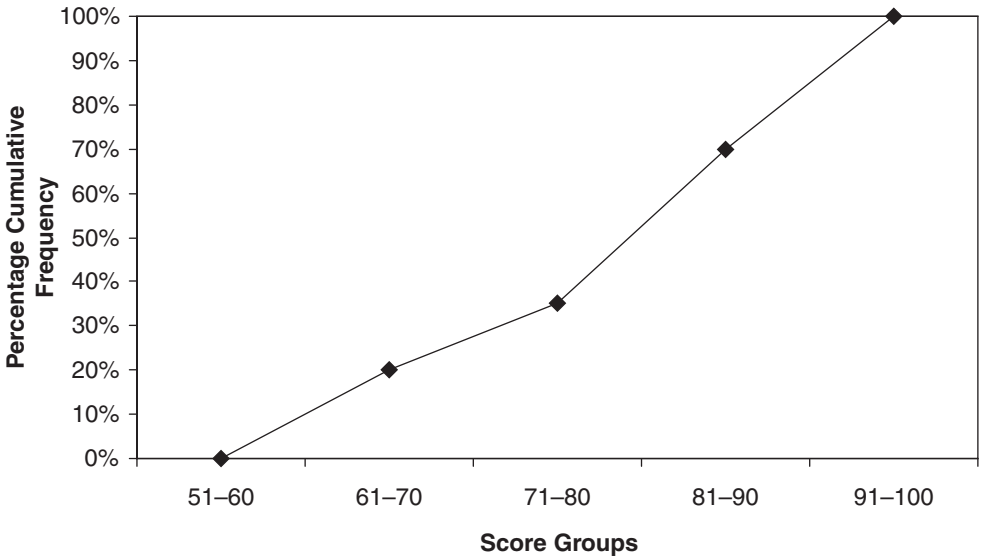



**Figure 2** Ogive of the Scores Obtained by Students on a Final Exam, Using Grouped Data

The relative cumulative frequencies are obtained by dividing each cumulative frequency by the total number of observations (the sum of all frequencies). The percentage cumulative frequencies are obtained by multiplying the relative cumulative frequency by 100. In relative cumulative frequency terms, it can be concluded from Table 3 that, for example, a proportion of 0.20 of students scored 70 points or less on the final exam. In percentage terms, 20% of the students scored 70 points or less.

Similarly to the ungrouped data case, the cumulative frequency of the highest score always equals the total number of observations (20 in this example). The corresponding relative cumulative frequency is always 1.00, and the corresponding percentage cumulative frequency is always 100%.

The ogive for this set of data (Figure 2) is based on any of the cumulative frequencies calculated in Table 3.

## Using Ogives to Obtain Percentiles

Percentiles are a measure of position. The $k$th percentile of a distribution, $P_k$, is a value such that $k$ percent of the observations are less than or equal to $P_k$, and $(100 - k)$ percent are greater than or equal to $P_k$. For example, the 70th percentile of the distribution of grades obtained by students at a final exam is a value such that 70% of the scores fall at or below it. The median (or the second quartile) is the 50th percentile, the value where 50% of the observations fall at or below it, and the other 50% fall at or above it. The other two commonly used percentiles are the first quartile and the third quartile. The first (or lower) quartile, Q1, is the 25th percentile; the third (or upper) quartile, Q3, is the 75th percentile.

The term often associated with percentiles in the literature is the *percentile rank*. Whereas the percentile refers to the value corresponding to a particular observation, the percentile rank refers to the percentage of observations at or below that value. Thus, if the 70th percentile of the distribution of grades obtained by students at a final exam is 89 points, then 70% of the scores fall at or below 89 points; the percentile is 89 points, whereas the percentile rank is 70%.

Percentile ranks provide the place of a particular score relative to all other scores above and below it. Thus, they show how a particular score compares to the specific group of scores in which it appears. For example, knowing that a student obtained 78 points on an exam, the percentile rank shows how well the student did compared to the

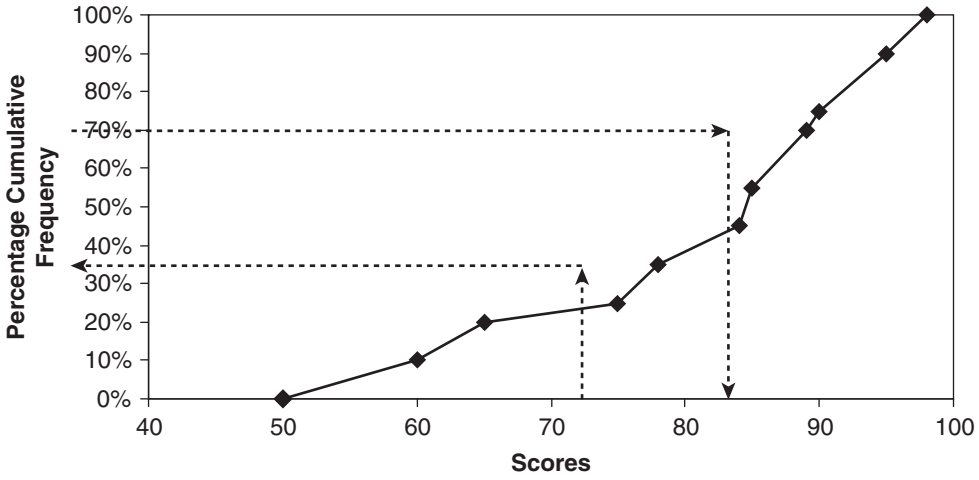



**Figure 3** Using the Ogive of the Scores Obtained by Students on a Final Exam to Calculate Percentiles

other students in his or her class. If the students obtained very good scores overall, a score of 78 may represent poor performance and thus correspond to a lower percentile. If the test was difficult and the scores were low overall, a score of 78 may actually be among the highest and thus correspond to a higher percentile. Percentiles and percentile ranks may also be used to compare a sample to another sample. Many times, one of the two samples compared is a normative (standard or control) sample.

One way to calculate percentiles and percentile ranks, and to compare samples, is by using the ogive corresponding to the analyzed set of data. However, the ogive must be drawn with very high precision for the results to be accurate. It is generally recommended that percentiles and/or percentile ranks be calculated using other methods (i.e., mathematical formulas), because the graphical approach may yield biased results, especially when a very precise representation of the corresponding ogive on grid paper is not available.

To obtain the 70th percentile, the point corresponding to 70% on the $y$-axis is first marked. A line is then drawn parallel to the $x$-axis until it intersects the ogive. From that point, a line is drawn parallel to the $y$-axis until it intersects the $x$-axis. This intersection point with the $x$-axis indicates the score corresponding to the 70th percentile: 89. Similarly, the score corresponding to the median (the 50th percentile) is 85 and the score corresponding to the upper quartile (Q3) is 90.

The percentile rank corresponding to a certain score may be obtained by reversing this process. For example, to obtain the percentile rank corresponding to 78 points, the position corresponding to 78 points on the $x$-axis is marked first; a vertical line is then drawn parallel to the $y$-axis until it reaches the ogive, and, from the intersection point, a horizontal line is drawn parallel to the $x$-axis until it reaches the $y$-axis. This intersection point with the $y$-axis indicates the percentile rank corresponding to a score of 78 points: 35%. This means that 35% of the students scored 78 points or less on the final exam.

The ogive may also be used to calculate the interquartile range. The interquartile range is a measure of spread showing the distance between the lower and the upper quartiles; it is often used to identify outliers. For the distribution graphed by the ogive in Figure 3, the lower and the upper quartiles may be calculated by employing the method described above. The values of the lower and upper quartiles are 75 and 90, respectively; consequently, the interquartile range is $90 - 75 = 15$.

The ogive may also be useful when comparing data from two or more samples or populations. For example, plotting the ogives for the scores obtained by two different groups of students on an exam may indicate which batch may have scored better. If the 25th percentile corresponding to Batch A is 78 and the 25th percentile

corresponding to Batch B is 55, the conclusion that Batch A scored better on the respective test may be drawn.

*Oana Pusa Mihaescu*

***See also*** Cumulative Frequency Distribution; Descriptive Statistics; Frequency Distribution; Frequency Table; Percentile Rank

**Further Readings**

Berenson, M. L., & Levine, D. M. (1996). *Basic business statistics: Concepts and applications* (6th ed.). Englewood Cliffs, NJ: Prentice Hall.
Dunn, D. S. (2001). *Statistics and data analysis for the behavioral sciences*. Boston: McGraw-Hill.
Heiman, G. W. (2000). *Basic statistics for the behavioral sciences*. Boston: Houghton Mifflin.
Moore, D. S., & McCabe, G. P. (2006). *Introduction to the practice of statistics*. New York: W. H. Freeman.
Nunnally, J. C. (1975). *Introduction to statistics for psychology and education*. New York: McGraw-Hill.
Sanders, D. H., & Smidt, R. K. (2000). *Statistics: A first course*. Boston: McGraw Hill.
Welkowitz, J., Ewen, R. B., & Cohen, J. (1976). *Introductory statistics for the behavioral sciences*. New York: Academic Press.

# Omega Squared

Omega squared ($\omega^2$) is a descriptive statistic used to quantify the strength of the relationship between a qualitative explanatory (independent or grouping) variable and a quantitative response (dependent or outcome) variable. The relationship is interpreted in terms of the proportion of variation in the response variable that is associated with the explanatory variable. As a proportion, it can have values between 0 and 1, with 0 indicating no relationship and 1 indicating that all of the variation in the response variable is attributed to the explanatory variable. Omega squared is used as an effect-size index to judge the meaningfulness of the observed relationship identified using the analysis of variance $F$ test. It can supplement the results of hypothesis tests comparing two or more population means. The research design may be either experimental, involving the random assignment of units to levels of the explanatory variable (e.g., different drug treatments), or nonexperimental, involving a comparison of several groups representing existing populations (e.g., underweight, normal weight, overweight, obese).

## A Data Example

Consider an experimental study designed to evaluate the merits of three drug treatments to reduce the number of cigarettes smoked (i.e., reduce smoking behavior). From a volunteer group of 150 moderate to heavy smokers, 50 individuals are randomly assigned to each of the three identified drug treatments ($n = 50$). After a 6-week treatment period, participants are asked to record the number of cigarettes smoked during Week 7. Hypothetical means and standard deviations are reported in Table 1.

To test the hypothesis that there is no difference in the average number of cigarettes smoked by individuals exposed to the three drug treatments (i.e., $H_0 : \mu_1 = \mu_2 = \mu_2 = \mu_3$), the analysis of variance (ANOVA) $F$ test could be used. The sample means estimate the population means, but they are subject to sampling error. The statistical test provides information on whether the observed difference among sample means provides sufficient evidence to conclude that population means differ, or whether it is just a reflection of sampling error. More importantly, the ANOVA $F$ test does not provide any indication of the amount of difference or the strength of the relationship between the treatments and smoking behavior. The statistical test provides only the probability of the observed $F$ statistic if the null hypothesis (i.e., the drugs are equally effective) is true. Differences among population means may be very small, even trivial, and these differences can be detected by the statistical test if the number of individuals in the samples is sufficiently large. Statistical significance does not

**Table 1** Mean Number of Cigarettes Smoked and Standard Deviations for Each of Three Drug Treatments

| *Treatment* | *Mean* | *Standard Deviation* |
|---|---|---|
| Drug 1 | 15 | 9.8 |
| Drug 2 | 18 | 10.7 |
| Drug 3 | 12 | 9.2 |

imply meaningful significance. For the data in Table 1, the results might be judged to be statistically significant [$F(2,147) = 4.573$, $p = .011$], and it might be concluded that the drug treatments are differentially effective with respect to smoking behavior. Or, it might be concluded that a relationship does exist between the drug treatments and smoking behavior.

## Effect Size

Omega squared provides a measure of association between the grouping factor (e.g., drug type) and the outcome variable (e.g., number of cigarettes smoked). That is, it estimates the proportion of variation in the outcome variable that is associated with the grouping variable. Theoretically, the variance of the population means, $\sigma^2_G$, plus the variance of individual scores within a population, $\sigma^2_{I|G}$ (which is assumed to be equal for all populations), determines the total variance in the response variable, $\sigma^2_Y$, (i.e., $\sigma^2_Y = \sigma^2_G + \sigma^2_{I|G}$). In a balanced design (equal number of units per group), $\omega^2$ is

$$\omega^2 = \frac{\sigma^2_G}{\sigma^2_Y} = \frac{\sigma^2_G}{\sigma^2_G + \sigma^2_{I|G}}.$$

To estimate $\omega^2$,

$$\hat{\omega}^2 = \frac{SS_{\text{between}} - (J-1)MS_{\text{within}}}{SS_{\text{total}} + MS_{\text{within}}},$$

where $SS_{\text{between}}$ is the between levels of the grouping variable Sum of Squares, $J$ is the number of groups, $SS_{\text{total}}$ is the total sum of squares, and $MS_{\text{within}}$ is the within-groups mean square and computed as the sum of squares within groups divided by the degrees of freedom within groups [$MS_{\text{within}} = \frac{SS_{\text{within}}}{(N-J)}$]. For the data in Table 1, $SS_{\text{between}} = 900$, $J = 3$, $SS_{\text{within}} = 14{,}463.33$, $MS_{\text{within}} = 98.39$, $SS_{\text{total}} = 15{,}363.33$, and $\hat{\omega}^2$ equals

$$\hat{\omega}^2 = \frac{900 - (3-1)98.39}{15,363.33 + 98.39} = .045.$$

These results can be interpreted as 4.5% of the variation in the number of cigarettes smoked is explained by the drug treatments. Whereas the maximum value possible for $\hat{\omega}^2$ is 1, actual results from empirical studies are much smaller. One difficulty with $\hat{\omega}^2$ is evaluating its magnitude. Many researchers adopt guidelines for interpreting omega squared suggested by Jacob Cohen to define small, medium, and large relationships as .01, .06, and .14, respectively. These guidelines were originally based on research findings reviewed by Cohen in the 1960s and 1970s. They were not intended to be standards for all response variables, with all populations, in all contexts. However, these guidelines have become standards for many researchers. The interpretation of omega squared should be based on literature reviews and findings reported by other researchers conducting similar inquiries with similar populations but using different response measures.

## Factorial Designs

Many research studies include more than a single explanatory variable. These additional explanatory variables may be manipulated variables, or they may be some characteristic of the participants. An example of a second manipulated variable for the study described above might be the time of day when the drugs are administered. Regardless of which drug is taken, it might be of interest to determine whether the time of day when a drug is administered has an effect on the number of cigarettes smoked. That is, a drug may be more effective if taken in the morning rather than at night. An example of a participant characteristic that might be included in the drug study is previous smoking behavior. That is, regardless of which drug is taken, can smoking behavior be affected more among moderate smokers than heavy smokers? The nature of the second factor in the study has an important influence in how omega squared is computed and interpreted. Table 2 presents hypothetical group means for a 2 × 3 factorial design. These data will be used to discuss how the computation and interpretation of $\hat{\omega}^2$ changes. We will assume the total sample size remains the same with 25 participants available for each Factor 2 by Drug combination ($n = 25$).

**Table 2** Hypothetical Group Means for a 2 × 3 Factorial Design Studying Drug Effectiveness

| | | *Drug 1* | *Drug 2* | *Drug 3* | *Row Means* |
|---|---|---|---|---|---|
| Factor 2 | $b_1$ | 10 | 13 | 7 | 10 |
| | $b_1$ | 20 | 23 | 17 | 20 |
| | *Column means* | 15 | 18 | 12 | |

## Partial Omega Squared

If Factor 2 is a manipulated factor (e.g., time of day), the addition of this factor to a study comparing drug treatments will increase the total variance in the data set. Consequently, the total sum of squares will increase but neither $SS_{\text{between}}$ nor $MS_{\text{within}}$ will be affected. [Note that for the new data set, there is no interaction ($SS_{\text{interaction}} = 0$) between Factor 2 and drug treatment.] The mean number of cigarettes smoked under each drug type remains the same, and adding a manipulated factor does not change the estimate of variance of individuals within treatment combinations. With the data in Table 2, $SS_{\text{total}} = 19{,}113.33$ and the sum of squares associated with Factor 2 is 3,750. Omega squared computed with Factor 2 in the design gives the following results:

$$\hat{\omega}^2 = \frac{900 - (3-1)98.39}{19,113.33 + 98.39} = .037.$$

Although the drug treatment means are identical in Tables 1 and 2, the estimate of the strength of the relationship between drug treatment and smoking behavior is reduced 17.7% to .037 because of the addition of Factor 2. What is needed is a statistic that is not influenced by the presence of other manipulated factors in the study.

Partial omega squared ($\omega_p^2$) estimates the relationship between a factor (a manipulated factor or interaction) and the response variable, ignoring all other manipulated factors and interactions. For a balanced factorial design, partial omega squared is defined as

$$\hat{\omega}_p = \frac{\sigma^2_{\text{factor}}}{\sigma^2_{\text{factor}} + \sigma^2_{\text{error}}},$$

where $\sigma^2_{\text{factor}}$ is the variance due to the factor of interest, which may be a main effect or an interaction, and $\sigma^2_{\text{error}}$ is the error variance for that effect (e.g., $MS_{\text{within}}$). Partial omega squared is estimated as

$$\hat{\omega}_p = \frac{SS_{\text{factor}} - df_{\text{factor}}MS_{\text{error}}}{SS_{\text{factor}} + (N - df_{\text{factor}})MS_{\text{error}}},$$

where, in a completely randomized factorial design, $SS_{\text{factor}} = SS_{\text{between}}$ and $MS_{\text{error}} = MS_{\text{within}}$. Using the data in Table 2, partial omega squared is computed as

$$\hat{\omega}_p^2 = \frac{900 - (3-1)98.39}{900 + (150-2)98.39} = .045.$$

Partial omega squared thus provides the same estimate of the relationship between drug treatments and the number of cigarettes smoked when time of day is included as a factor in the design and when drug treatments is the only factor in the study.

When the second factor in the study is a characteristic (C) of the participants (e.g., moderate or heavy smoker), the use of partial omega squared will result in the estimation of the relationship between the manipulated factor (e.g., drug treatments) and the response variable that is much larger than the estimate obtained when the characteristic is not included. A factor that is a characteristic of the participants does not increase the total variation in the response variable like a manipulated variable does. Rather, the variation associated with the characteristic *is* part of the within-group variance. That is, $\sigma^2_{\text{within}} = \sigma^2_C + \sigma^2_{I|G}$. Keeping the total sum of squares the same for Tables 1 and 2 (i.e., $SS_{\text{total}} = 15363.33$), the $MS^*_{\text{within}}$ for the data in Table 2 would equal

$$74.398\left[\frac{SS_{\text{within}} - SS_C}{N - JK} = \frac{14,463.33 - 3750}{150 - (2)(3)}\right].$$

Using the formula for partial omega squared, the relationship between drug treatments and smoking behavior would equal

$$\hat{\omega}_p^2 = \frac{900 - (3-1)74.398}{900 + (150-2)74.398} = .063.$$

When the second factor is a characteristic of the participants, the within-group variance is reduced and partial omega squared provides an estimate of the relationship between the first manipulated factor and the response variable that is much greater than an estimate of the relationship when the characteristic is not included in the analysis. In the present example, partial omega squared is 40% greater than when previous smoking behavior is not considered.

To obtain consistent estimates of the relationship between an explanatory and a response variable across a variety of research designs, serious consideration must be given to the design features of the studies. For a between-group design where all of the factors being studied are manipulated, partial omega squared is appropriate, but in the case of one manipulated factor and one factor being a characteristic of the participants, omega squared should be computed for the manipulated factor.

## Limitations

In addition to the difficulty of judging the magnitude of $\hat{\omega}^2$ in terms of the meaningfulness of the observed relationship, this effect-size index has several limitations that must be taken into consideration when interpreting its value. Another limitation is the instability of the statistic. A confidence interval (CI) for the parameter can be very wide. For example, using the data in Table 1, $\hat{\omega}^2 = .045$ and the .95CI is (.003, .137). Using Cohen's guidelines to interpret $\hat{\omega}^2$, the current estimate of the relationship between smoking behavior and the drug treatments would be judged to be from virtually nonexistent to large. Given that the present estimate was obtained on a sample size that is much larger than is often found in experimental research, it should be clear from this example that researchers must be careful not to overinterpret the magnitude of the point estimate for $\omega^2$.

A second limitation associated with the interpretation of $\hat{\omega}^2$ is its dependence on the levels and strength of the treatment variable. In the present example, three drugs are evaluated; if a control (placebo) condition had been included, or if greater differences in drug doses were included, it is likely that the numerical value for $\hat{\omega}^2$ would increase. The specific levels of the explanatory variable chosen for the study (e.g., types of drugs) influence the magnitude of omega squared.

Another limitation with the numerical value of $\hat{\omega}^2$ is that it is affected by the reliability of both the treatments and the outcome measure. Inconsistency with the implementation of the treatments or how the outcome variable is measured will reduce the numerical value of $\hat{\omega}^2$. Furthermore, the heterogeneity of the population studied affects $\hat{\omega}^2$. The greater the variability among the units being studied, the lower the numerical value for omega squared. In addition, omega squared was developed for a fixed-effect analysis of variance (ANOVA) model. ANOVA assumes that the populations being compared have variances that are identical. Omega squared makes the same assumption. If sample sizes are unequal, violating this assumption is much more serious.

Finally, the research design used in the inquiry can affect the numerical value of $\hat{\omega}^2$. The use of covariates, blocking variables, or repeated measures can make estimates of $\hat{\omega}^2$ incompatible with estimates of omega squared from completely randomized studies. Repeated measures, blocking variables, and covariates reduce error variance, resulting in greater statistical power for hypothesis tests for comparing population means and greater values for omega squared. But the populations being compared are not the same as those compared in a completely randomized study, and so values of omega squared are not comparable.

## Conclusion

Omega squared can be a useful descriptive statistic to quantify the strength of the relationship between a qualitative explanatory variable and a quantitative response variable. It has the desirable feature of being a proportion and therefore has values between 0 and 1. Because statistical significance is influenced by the number of units in the study, virtually any relationship can be judged to be statistically significant if the sample size is large enough. Omega squared is independent of sample size, so it provides an index to judge the practicality or meaningfulness of an observed relationship.

Omega squared does have several limitations that should be considered when interpreting its value. First, there is no universally agreed-upon criterion for judging the importance of the estimated strength of relationship. Second, omega

squared is an unstable statistic having a wide confidence interval. Third, the reliability of the response and explanatory variables influences the magnitude of the estimated relationship. Fourth, the choice of levels of the explanatory variable affects omega squared. Fifth, the variability and equality of the populations studied can influence its value. And sixth, design features can affect the numerical value of the estimated relationship.

In spite of these limitations, omega squared can provide useful information to aid the interpretation of study results and should be reported along with the results of statistical hypothesis tests.

*Stephen Olejnik*

***See also*** Between-Subjects Design; Block Design; Effect Size, Measures of; Eta-Squared; Experimental Design; Partial Eta-Squared; *p* Value; Single-Subject Design

**Further Readings**

Cohen, J. (1988) *Statistical power analysis for the behavioral sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Grissom, R. J., & Kim, J. J. (2001). Review of assumptions and problems in the appropriate conceptualization of effect size. *Psychological Methods, 6,* 135–146.

Keppel, G., & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook* (4th ed.). Englewood Cliffs, NJ: Prentice Hall.

Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology, 25,* 241–286.

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods, 8,* 434–447.

Steiger, J. H. (2004). Beyond the *F* test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods, 9,* 164–182.

# Omnibus Tests

Omnibus tests are statistical tests that are designed to detect any of a broad range of departures from a specific null hypothesis. For example, one might want to test that a random sample came from a population distributed as normal with unspecified mean and variance. A successful *genuine* omnibus test would lead one to reject this hypothesis if the data came from any other distribution. By contrast, a test of normality that was sensitive specifically to thick-tailed distributions such as Cauchy would not be considered an omnibus test. A genuine omnibus test is consistent for *any* departure from the null, rejecting a false null hypothesis with probability approaching unity as the sample size increases; however, some statisticians include in the omnibus category certain broad-spectrum tests that do not meet this stringent condition. This entry presents general types, general features, and examples of omnibus tests.

## Categories

Here are some general types of omnibus tests: (a) tests for differences in means of groups of data distinguished by "treatments," such as in one-way analysis of variance (ANOVA); (b) tests that combine the results of independent tests of various hypotheses to indicate whether one or more of them is false; (c) tests for time-series models, such as portmanteau tests for adequacy of autoregressive-integrated-moving-average (ARIMA) models and tests for structural breaks; (d) goodness-of-fit tests for hypotheses about marginal or joint distributions, such as tests for univariate or multivariate normality.

## General Features

All of these procedures have the advantage of detecting whether *all* aspects of the data are consistent with the hypothesis in question, but this breadth of scope comes at a price: (a) Omnibus tests typically do not pinpoint the specific features of the data that are most at variance with the null; and (b) omnibus tests typically have less power against specific departures from the null hypothesis than do tests with narrower focus. Because of these considerations, conducting an omnibus test is most often just a *first* step in statistical analysis. It can help in deciding whether further study is appropriate; for example, if an ANOVA test rejects that all treatment means are equal, one might then seek to identify and explain specific sources of

difference. An omnibus test can also help in selecting the best *methods* for further analysis; for example, if the data are consistent with the normality hypothesis, then one has available the usual minimum-variance estimators and uniformly most powerful tests for mean and variance. However, in applying a sequence of tests to *given* data, there is the danger of data snooping (performing repeated analytical tests on a data set in the hope of finding a significant test). Thus, when applied *unconditionally,* a valid .05-level test that two specific treatment means differ would reject a true "no-difference" null in 5% of repeated samples; yet the same test would reject more often if applied to the two of $m > 2$ treatments with the most divergent means in samples for which ANOVA rejects the full $m$-variate null.

## Examples

### *One-Way ANOVA*

Given independent samples (vectors) $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$ of sizes $n_1, \ldots, n_m$, we test $H_0$ that all $m$ population means are equal, assuming that the populations are normal with the same variance. $H_0$ is rejected at level $\alpha$ if $\Sigma_j n_j (M_j - M_N)^2 / [(m-1)S_p^2]$ exceeds the upper-$\alpha$ quantile of the $F$ distribution with $m - 1$ and $N = \Sigma_j n_j$ degrees of freedom, where $M_j$ is the mean of sample $j$, $M_N$ is the overall mean, and $S_p^2$ is the weighted average of sample variances.

### *Fisher's Test*

Given probability values ($p$ values) $P_1, \ldots, P_m$ from *independent* tests of $m$ null hypotheses, the test rejects at level $\alpha$ that *all* are true if $-2[\ln(P_1) + \cdots + \ln(P_m)]$ exceeds $\chi^2_\alpha(m)$, the upper-$\alpha$ quantile of $\chi^2(m)$. The idea is that the combined results might signal statistical significance, although no individual test does so.

### *Time Series Tests*

Time series $Y_1, \ldots, Y_{T+d}$ is modeled as an ARIMA($p$,$d$,$q$) process, and residuals are calculated as $u_t = Z_t - \phi_1 Z_{t-1} - \cdots - \phi_p Z_{t-p} + \theta_1 u_{t-1} + \cdots + \theta_q u_t \theta_{t-q}$, where $Z_t = \Delta^d Y_t$ is the $d$th-order difference. With $r_1, \ldots, r_m$ as the residuals' autocorrelations at lags $1, 2, \ldots, m$, the Box–Pierce portmanteau test of the model's adequacy rejects at level $\alpha$ if $T\Sigma r_j^2 \geq \chi^2{}_\alpha(m)$. If residuals $\{e_t\}$ from structural model $Y_t = \mathbf{X}_t\boldsymbol{\beta} + \varepsilon_t$ are calculated *recursively* using at each $t$ only prior data, then a structural break—a change in one or more coefficients at any $t$—can be detected with a cumulative sum test involving sums of squares of the $\{e_t\}$.

### *Tests of Fit*

The many omnibus tests for whether data conform to some particular family of distributions include those based on (a) relative frequencies, such as Pearson's chi-square; (b) moments, such as skewness and kurtosis tests; (c) order statistics; (d) empirical distribution functions; and (e) generating functions. Only tests in Groups (c)–(e) can count as *genuine* omnibus tests, as defined in the opening paragraph. In Group (c), the best known are the Shapiro–Wilk and Shapiro–Francia tests for normality. These compare estimates of standard deviation based on order statistics (the observations in rank order) and on sample moments. In Group (d) are the Kolmogorov–Smirnov (KS) and Anderson–Darling (AD) tests. KS is based on the maximum difference between the cumulative distribution functions of sample and model, $sup|F_n(x) - F_0(x)|$. AD is based on an integral of squared differences, $\int [F_n(x) - F_0(x)]^2 W(x) \cdot dx$, where $W$ is a weight function. This has a simple computational form in terms of order statistics and is typically far more powerful than KS. Group (e) includes tests that compare model and sample moment-generating functions, probability-generating functions, and characteristic functions (c.f.s). With $\phi_0(t) = E(e^{itX}|H_0)$ and $\phi_n(t) = n^{-1} \Sigma_j \exp(itX_j)$ as the respective c.f.s, tests based on $\int [\phi_n(t) - \phi_0(t)]^2 W(t) \cdot dt$ with appropriate $W$ are genuine omnibus tests for any distributional hypothesis.

*Thomas W. Epps*

***See also*** Analysis of Variance (ANOVA); Hypothesis; Kolmogorov–Smirnov Test; $p$ Value

### Further Readings

Casella, G., & Berger, R. L. (1990). *Statistical inference.* Belmont, CA: Duxbury.

D'Agostino, R. B., Stephens, M. A. (1986). *Goodness-of-fit techniques*. New York: Marcel Dekker.
Epps, T. W. (2005). Tests for location-scale families based on the empirical characteristic function. *Metrika, 62,* 99–114.
Greene, W. H. (1997). *Econometric analysis*. Upper Saddle River, NJ: Prentice Hall.

# One-Tailed Test

One-tailed test is a method of hypothesis testing where the alternative hypothesis specifies in which direction the parameter differs from the value stated in the null hypothesis. That is, the alternative hypothesis states if the parameter is above or below the value in the null hypothesis. One-tailed hypothesis testing is widely used in quantitative research when the direction of the population parameter's deviation from the value in the null hypothesis can be predicted in advance or when researchers are interested in results in a specific direction. This entry explains one-tailed tests in connection to other aspects of hypothesis testing and describes contexts in which one-tailed tests are the appropriate type of hypothesis testing.

## Alternative Hypotheses: Directional Versus Nondirectional

In hypothesis testing, null hypotheses ($H_0$) are tested against statistical alternative hypotheses ($H_a$). Alternative hypotheses can be set up as nondirectional or directional. A nondirectional $H_a$ states that the parameter differs from the value in the null hypothesis with no indication of the direction of the difference. For example, given $H_0$ stating that the population mean for reading achievement is 100 ($H_0 : \mu = 100$), the nondirectional $H_a$ states that the population mean is different from 100 ($H_a : \mu \neq 100$). Thus, the nondirectional $H_a$ does not specify if the population mean is greater or less than 100. On the other hand, a directional $H_a$ not only states that the parameter deviates from the value in the null hypothesis, but also specifies the direction of the deviation. For the aforementioned $H_0 : \mu = 100$, a directional $H_a$ can be that the population mean is either greater than 100 ($H_a : \mu > 100$) or less than 100 ($H_a : \mu < 100$).

The type of hypothesis testing in which $H_0$ is tested against a nondirectional $H_a$ is called a two-tailed test, whereas the one in which $H_0$ is tested against a directional $H_a$ is called a one-tailed test. The procedures for conducting one- or two-tailed tests are fundamentally similar. The difference between one- and two-tailed tests lies in the location of the region of rejection in sampling distributions.

Each hypothesis testing, whether one-tailed or two-tailed, starts with setting up the null and alternative hypotheses before collecting data. Thus, researchers need to decide whether they will conduct a one-tailed or a two-tailed test before data collection. The second step in hypothesis testing that takes place before data collection is deciding the level of significance, alpha value. The level of significance is the probability of rejecting $H_0$ when it is actually true (i.e., the probability of Type I error). As with other probability values in hypothesis testing, alpha is obtained from sampling distributions. For a statistic of interest (e.g., $\overline{X}$ or *SD*), a sampling distribution is a distribution of an infinite number of sample statistics where the samples are the same size random samples as the actual sample from a population where $H_0$ is true. In a sampling distribution, the proportion of the area covered by any range of sample statistics represents the probability of that range of sample statistics obtained for a random sample from a population where $H_0$ is true. Accordingly, the alpha value represents the proportion of the area in a sampling distribution that is equal to the probability of Type I error. For example, for a sample of size 50, if the probability of Type I error is 5%, in the sampling distribution, which consists of means for every random sample of size 50, the area covered by the range of values that correspond to Type I error is 5% of the whole area. The area represented by alpha is also referred to as the *region of rejection.*

## Region of Rejection

After data collection, a sample statistic is calculated for the sample in hand. The hypothesis testing proceeds to obtain the probability of the observed statistic or a more extreme one on a random sample from a population where $H_0$ is correct, which is called the $p$ value. If the $p$ value is

less than the alpha level, then $H_0$ is rejected in favor of $H_a$. If the $p$ value is larger than the alpha, $H_0$ fails to be rejected.

The sample statistic values for which the null hypothesis is rejected lie in the region of rejection. The region of rejection consists of sample statistics that are highly unlikely if the null hypothesis were true. In other words, the region of rejection consists of sample statistics that are extreme values of the alternative hypothesis. Thus, the region of rejection is directly related to the alternative hypothesis. For example, in a one-tailed test where $H_0$ is $\mu = 100$ and $H_a$ is $\mu > 100$, the region of rejection is located in the right tail of the sampling distribution. The sample statistics in the region of rejection are larger than 100 sufficiently enough to be considered highly unlikely at .05 level of significance if observed in a sample belonging to a population where $H_0$ is true. In the case of a one-tailed test where $H_0$ is $\mu = 100$ and $H_a$ is $\mu < 100$, the region of rejection is located in the left tail of the test statistic sampling distribution. The sample statistics in the region of rejection are smaller than 100 sufficiently enough to be considered highly unlikely at .05 level of significance if observed in a sample belonging to a population where $H_0$ is true.

The area of the region of rejection is determined by the alpha level that is set at the onset of a study. For example, if the alpha level is set at .05, the region of rejection covers 5% of the total area in the sampling distribution.

## Example

A heuristic example is used to explain the procedures for a one-tailed test. Although the sampling distribution and the sample statistics will differ, steps used in this example for conducting a one-tailed test can be generalized to other hypothesis-testing situations. This example compares male and female students' technology acceptance levels. The first step in one-tailed hypothesis testing is setting up $H_0$ and $H_a$ even before data collection:

$H_0$ : The population means are the same ($\mu_{\text{female}} - \mu_{\text{male}} = 0$).

$H_a$ : The mean for females is greater than the mean for males ($\mu_{\text{female}} > \mu_{\text{male}}$).

Because the $H_a$ states that $\mu_{\text{female}}$ is *larger* than $\mu_{\text{male}}$, we are interested in the right-hand tail of the sampling distribution. In other words, the critical region for rejecting the null hypothesis lies in one (i.e., right) tail of the distribution. In one-tailed tests, the direction of the inequality in $H_a$ determines the location of the region of rejection. If the inequality is $>$ in $H_a$, the region of rejection is in the right tail of the sampling distribution. If the inequality is $<$ in $H_a$, the region of rejection is in the left tail of the sampling distribution.

The second step in hypothesis testing involves deciding the alpha value as the criterion for rejecting $H_0$. In this example, alpha was set at .05. The region of rejection, which constitutes sample statistics that are highly unlikely if the null hypothesis were true, covers 5% of the total area of sampling distribution and lies in the right tail.

Then, researchers collect data on 20 female and 20 male students. In these samples, the mean technology acceptance levels for females and males were 100 and 70, respectively. Thus, the difference in sample means is $100 - 70 = 30$. It is important to note that when obtaining the difference in the sample means, we subtracted $\mu_{\text{male}}$ from $\mu_{\text{female}}$, keeping the order in $H_a$. Hypothesis testing proceeds to evaluate if a difference of 30 in the sample means is large enough to reject the $H_0$ that the means do not differ in the population. Therefore, the probability (i.e., $p_{\text{calculated}}$) of obtaining such a difference or a more extreme one in a population where $H_0$ is true will be obtained.

To determine the $p_{\text{calculated}}$, the sampling distribution of the differences between two sample means will be used. With the sampling distribution, first a standardized score that quantifies the difference between the sample statistic and the hypothesized population parameter (i.e., test statistic) will be obtained. In this example, the sampling distribution of mean differences is derived assuming the mean difference in the populations is zero, which was stated in the null hypothesis as $\mu_{\text{female}} - \mu_{\text{male}} = 0$. The test statistic that will be used to compare the mean

differences is the $t$ statistic. A general formula for the $t$ statistic is

$$t \text{ statistic} = \frac{\text{Sample statistic} - \text{Hypothesized parameter}}{\text{Standard error of the statistics}}.$$

For this example, assuming the standard error of the mean differences is 10, the $t$ statistic is as follows:

$$t = \frac{(100 - 70) - 0}{10} = 2.0.$$

We can interpret a $t$ statistic of 2.0 as a mean difference of 30 in the sample is 2.0 standard deviations away from the hypothesized population mean difference, which is zero.

In the last step of the hypothesis testing, rese-archers can determine the $p$ value, which is the probability of obtaining a $t$ statistic of 2.0 or more in a population where $H_0$ is true. Because, in this example, a one-sided $H_a$ is used, only the probability of $t \geq 2.0$ is needed. Using a commonly available $t$-statistic sampling distribution, the prob-ability of $t \geq 2.0$ is approximately .02625, which is smaller than the alpha value of .05. Because the $p$ value is smaller than the alpha, the decision about $H_0$ is "reject."

### One-Tailed Versus Two-Tailed Tests

In the above example, if a two-tailed hypothesis testing were conducted, the alternative hypothesis would be $\mu_{\text{female}} \neq \mu_{\text{male}}$. Because in a two-tailed test, the alternative hypothesis does not have a direction, when researchers determine $p_{\text{calculated}}$, they consider not only obtaining such a difference or *more* in a population where $H_0$ is true but also obtaining such a difference or *less* in a population where $H_0$ is true. Thus, when a two-tailed $H_a$ is used, both the probability of $t \geq 2.0$ and $t \leq 2.0$ are needed. Because $t$-distributions are symmetrical, the probability of $t \leq 2.0$ is approximately .02625 also. Therefore, the probability of $t \leq 2.0$ or $t \geq 2.0$ (i.e., $p$ value) is around $.02625 + .02625 = .052$, which is larger than the alpha value of .05. Because the $p$ value is larger than the alpha in the two-tailed case, the decision about $H_0$ is "fail to reject."

As seen in the above example, the decision about $H_0$ may differ when a two-tailed rather than a one-tailed test is used. One-tailed tests provide more power to detect effects in one direction than do two-tailed tests. The increased power of one-tailed tests is due to the fact that the rejection region for a one-tailed test is located in only one of the tails of the sampling distribution. In other words, the probability of an effect in the opposite direction is ignored. Thus, the $p_{\text{calculated}}$ in a one-tailed test is smaller than the $p_{\text{calculated}}$ for a two-tailed test (e.g., is a symmetric sampling distribution, the $p_{\text{calculated}}$ in a one-tailed test is half as much as the $p_{\text{calculated}}$ in a two-tailed test). If researchers have theoretical bases to expect a difference in only one direction, one-tailed tests would provide them with more power. However, if researchers lack such a theoretical base, they should be careful about not using one-tailed tests just because one-tailed tests provide more power. Because one-tailed tests are more likely to provide statistically significant results, one-tailed tests without a theoretical base should be taken with a grain of salt. Another cautionary remark about using one-tailed tests in the absence of a theoretical base is that if sample statistics yield an effect in the opposite direction from the one-sided $H_a$, the null hypothesis will not be rejected.

*Z. Ebrar Yetkiner and Serkan Ozel*

***See also*** Hypothesis; $p$ Value; Sampling Distributions; Significance Level, Concept of; Significance Level, Interpretation and Construction; Two-Tailed Test

#### Further Readings

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2003). *Applied statistics for the behavioral sciences* (5th ed.). Boston: Houghton Mifflin.

Thompson, B. (2006). *Foundations of behavioral sciences: An insight-based approach*. New York: Guilford.

## "On the Theory of Scales of Measurement"

The 1946 article by Stanley Smith Stevens titled "On the Theory of Scales of Measurement" defined measurement as the assignment of numerals to objects or events according to rules. Stevens

went on to discuss how different rules for assigning such numbers resulted in different types of scales. In the remainder of the article, Stevens elaborated on four such scales in terms of the mathematical transformations that can be conducted without changing their properties and the statistical operations that he considered permissible for each. Stevens's definition of measurement and the four levels of measurement he described have since become iconic in social science measurement.

The four scales of measurement specified by Stevens were the nominal, ordinal, interval, and ratio scales. The scales form a specific hierarchy from low (nominal) to high (ratio) that is ordered on the basis of the types of operations that must be supported in order for a measurement to achieve a particular level. For example, nominal scales must provide for a means of determining distinctiveness, whereas ordinal scales must support the determination of order. The operations are cumulative in the sense that a scale at a higher level supports all of the operations of the scales beneath it, while adding an additional operation. Because of the hierarchical nature of the four measurement scales, they are often referred to as the *levels of measurement*.

## Nominal Scales

The nominal scale is at the lowest level of Stevens's hierarchy. Scales at this level assign numbers only as labels that can be used to distinguish whether attributes of different objects are the same. For example, the numbers "1" and "2" might be assigned to males and females as a shorthand means of differentiating the two. Here, the numbers used are not meaningful in any numerical sense, but are simply chosen for convenience. Other numbers, such as 100 and 200, would serve equally well. Permissible transformations of numbers on a nominal scale include any one-to-one substitution. As an example, assigning the numbers 3 and 4 instead of 1 and 2 would not alter our ability to distinguish between males and females, as long as we knew the "rule" governing this assignment. Because the numbers assigned have no numerical meaning, the only statistics considered to be permissible are those based on counts of the number in each category. Thus, we could determine how many males and females there were, or whether males or females were more numerous. However, we have no way of determining whether one gender is larger or greater than another because numbers in the nominal scale are arbitrary and do not support such a numerical meaning.

## Ordinal Scales

In addition to the property of distinctiveness, ordinal scales must have the property of order, or of determining whether one object has more or less of an attribute than another. Thus, ordinal scales provide a means for ordering objects along a continuum of some sort. One familiar example is the outcome of a race. Contestants are ranked according to their order in crossing the finish line, with the first finisher regarded as faster than the second, and so on. Note that such ranks do not provide any information regarding the magnitude of difference between those with different ranks. In other words, based on an ordinal scaling, we have no way of knowing how *much* faster the first-place winner was than the second. Any transformation that preserves the original order is permissible for ordinal scales. For example, a constant such as two could be added to each rank, or different constants could be added to different ranks, as long as doing so did not change the original ordering. Statistical operations considered permissible for ordinal data include computation of the median, percentiles (although see Stevens, 1946, p. 679), and semi-interquartile range. Computations of statistics such as the mean or standard deviation are not considered permissible because they treat the intervals between scale points as equally spaced, which is not a property of ordinal scales.

Likert-type scales that use a response format such as *disagree* to *agree* are, strictly speaking, at the ordinal level of measurement because it is not clear that the intervals are equal. Therefore, statistics such as the mean and standard deviation, and statistical tests such as the $t$ test and $F$ test, are not considered appropriate. However, methodologists disagree about this, and many feel that the use of such statistics is not problematic provided that the scale has at least five response options that are approximately normally distributed.

### Interval Scales

The interval scale adds the requirement that the intervals between adjacent scale points must be equal. A common example is temperature as measured by the Fahrenheit or Centigrade scales, in which the difference between 50° and 51° is considered to be the same as the difference between 90° and 91°. Linear transformations of the form "$y = a + b * x$" are allowable for interval scales because they preserve the equal interval property. Nearly all of the parametric statistical operations, such as calculation of the mean, standard deviation, and product-moment correlation, and statistical tests such as the *t* test and *F* test, are considered permissible for intervally scaled data.

### Ratio Scales

Ratio scales add the property of an absolute zero point. By "absolute zero," Stevens meant a point indicating an absolute lack of the property being measured. For example, on a monetary scale, zero means the absolute lack of any money. Ratio scales are more commonly found in the physical than the social sciences because the definition of an absolute zero point is often problematic, and some argue unnecessary, in the latter. Ratio scales support all types of statistical operations. Numbers on a ratio scale can be legitimately transformed only through multiplication of each value on the scale by a constant. Adding a constant to each scale value is not permissible because this would change the value of the zero point, rendering it nonabsolute.

*Deborah L. Bandalos*

***See also*** Descriptive Statistics; Interval Scale; Levels of Measurement; Likert Scaling; Nominal Scale; Ordinal Scale; Parametric Statistics; Psychometrics; Ratio Scale

#### Further Readings

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 102,* 677–680.

Stevens, S. S. (1951). Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1–49). New York: Wiley.

## Order Effects

In presenting information to individuals or groups, or in assessing how individuals or groups feel or think about an issue, researchers face the possibility that the order of materials might have a significant influence on the behavior, decisions, or evaluations of interest. A number of research streams have developed over the years to address these issues, known as order effects, as they are of interest for both practical (e.g., the desire to reduce bias in responses) and theoretical reasons (e.g., testing theoretical notions about the stability or strength of attitudes).

In persuasion research, message and argument order effects refer to changes in the influence of messages and arguments on attitude when the order of messages or argument presentation is varied. In survey research, two types of order effects are often observed: question order effects and response order effects. A question order effect refers to cases where exposure to or answering an earlier question affects the response to a later question. A response order effect refers to cases where the probability of an option being chosen changes depending on where it is placed in the list of options.

### Message and Argument Order Effects in Persuasion Research

In the classic message order effect research by Carl Iver Hovland and his colleagues, primacy effects were said to be observed when the first of two opposing messages carried the greatest weight in some final judgment. Recency effects were said to be observed when the second of two opposing messages carried the greatest weight in a final judgment. In another line of research, Hovland and colleagues examined the persuasive influence of the order of information in a single persuasive message.

Most recent findings and reviews of the literature suggest that the motivation level with which an individual processes a first message is an important consideration. In general, primacy effects are more likely to be observed when a first message is processed under high levels of motivation and

elaboration (i.e., when the person forms a strong attitude when exposed to the first message). Recency effects are more likely when the first message is not processed extensively. Specific task instructions can reduce the likelihood of primacy effects (e.g., considering the information in chunks rather than piecemeal).

The line of research regarding argument order effects within a single message, along with the message order effect research, has focused on the persuasive outcome of exposure to materials. General results show that arguments presented first tend to have greater impact, especially when the list of arguments is presented verbally. More recent research suggests that individuals hold expectations about the order in which arguments are presented. Arguments are more effective when they are presented at positions where most important arguments are expected.

## Question and Response Order Effects in Survey Research

Surveys are conducted to obtain measures of attitude, behavior, and/or memories of past behaviors that are as accurate representations of reality as possible. Two types of order effects are often encountered in survey research. One type is observed when answers vary when question order is changed, and the other type is observed when the order of response alternatives leads to different choices by respondents. Both question order effects and response order effects challenge the validity of survey data and generate intense interest in systematic research into the underlying causes of such effects.

One of the earliest examples of question order effect came from a survey on Americans' opinion on allowing U.S. citizens to join the German versus British or French army prior to America's involvement in World War II. Americans were more receptive of allowing U.S. citizens to join the German army if the question was asked after they had answered the same question on allowing U.S. citizens to join the British or French army. In another example, survey respondents reported their views on whether or not it should be possible for women to obtain legal abortion under two scenarios: when there is strong chance for birth defect (birth defect item) and when the woman is married but does not want to have more children (married-woman item). When the birth defect item comes first, there was less support for pro-choice on the married woman item.

Since the 1950s, survey researchers have documented many examples in which question order altered responses to the target question. Reviews of the literature suggest that a variety of social, cognitive, and motivational processes affect how respondents comprehend the survey question, retrieve relevant information, form judgments, map judgments to response categories, and edit responses before giving out answers.

First, surveys can be viewed as structured conversations between researchers and survey participants. As with ordinary conversations, speakers are supposed to be truthful, relevant, informative, and clear. Unless instructed otherwise, survey respondents will likely follow these conversational norms in interpreting the meaning of questions and deciding what information is sought by the researcher. Thus, they may decide to exclude information already provided through earlier questions from their judgment in answering later questions in an attempt to be informative. Changes in question order can lead to the exclusion of different information from later judgment, thus producing variations in response.

Second, survey respondents use a variety of information to infer the intended meaning of questions and forming judgments. Previous questions can serve as a frame of reference and affect respondents' interpretation of later questions. Answering preceding questions may also make related information more accessible in memory and more likely to be used in forming judgments in later questions. Answering questions can also produce changes in mood and perceptions of how easy or difficult it is to recall relevant information. Such changes in subjective experience can confound reported judgments. In addition, answering a question may increase the accessibility of certain procedural knowledge or decision strategies that may then be used in answering later questions.

Third, answers to a previous question may also trigger differing degrees of motivational processes. In the previous example on granting U.S. citizens permission to enlist in the German army, the majority of survey respondents felt it was okay to

let U.S. citizens join the friendly armies of Britain and France. Having expressed this sentiment, they would feel compelled to grant similar permission to those who wish to join the German army because denying them would be perceived as a violation of the social norm of fairness. When asked outright about their opinion to grant permission to individuals seeking to enlist in the German army, survey respondents were unlikely to be affected by the motivation to be fair; thus, their responses tended to be less positive.

Most individuals are motivated to hold beliefs that are cognitively consistent and prefer others to see them as being consistent in their ideas and opinions. Answers to previous questions may commit a respondent to a particular position and create pressure to appear consistent when answering later questions.

In addition to question order effects, response order effects are also frequently observed in mail and telephone surveys. In mail surveys, survey participants tend to choose the option at the beginning of the list with greater frequency than the option at the end of the list. In telephone surveys, when the options are presented verbally, participants tend to choose the option at the end of list with greater frequency. Such differences have been attributed to limitations of working memory capacity and a general tendency of respondents to choose an acceptable answer instead of the most appropriate answer.

As the brief review above suggests, the order in which messages, arguments, questions, or options for answers are presented can have significant influence on the data obtained from participants. It is thus important for social scientists to consider and study the influence of such variables in their own applications.

*Curtis P. Haugtvedt and Kaiya Liu*

***See also*** Experimental Design; Experimenter Expectancy Effect; Laboratory Experiments; Nonclassical Experimenter Effects

**Further Readings**

Haugtvedt, C. P., & Wegener, D. T. (1994). Message order effects in persuasion: An attitude strength perspective. *Journal of Consumer Research, 21*(1), 205–218.

Hovland, C. I. (1957). *The order of presentation in persuasion*. New Haven, CT: Yale University Press.

Igou, E., & Bless, H. (2003). Inferring the importance of arguments: Order effects and conversational rules. *Journal of Experimental Social Psychology, 39*(1), 91–99.

Krosnick, J., & Alwin, D. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly, 51*(2), 201.

Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. New York: Academic Press.

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology*. San Francisco: Jossey-Bass.

Tourangeau, R., & Rasinski, K. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin, 103*(3), 299–314.

Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2000). *The psychology of survey response*. New York: Cambridge University Press.

## Ordinal Scale

In the hierarchy of measurement levels, the ordinal scale is usually considered the second lowest classification order, falling between the nominal and interval scales. An ordinal scale is a measurement scale that allocates values to variables based on their relative ranking with respect to one another in a given data set. Ordinal-level measurements indicate a logical hierarchy among the variables and provide information on whether something being measured varies in degree, but does not specifically quantify the magnitude between successive ranks. The measurement taxonomy, including the ordinal scale terminology, was first brought forth by psychologist Stanley Smith Stevens in his 1946 seminal paper and has subsequently guided the selection of appropriate statistical techniques despite debates on its limitations, data exceptions, and contemporary relevance.

Ordinal-scale variables may be further classified as strongly ordered or weakly ordered (as presented by J. Chapman McGrew, Jr., and Charles Monroe) depending on the ranking scheme. A strongly ordered variable is one in which the measurements

are continuous, sequentially ordered, and not strictly categorically dependent. For example, a list of the 10 most populous countries in the world receive rankings based on their relative population size to one another, demonstrating an order then from highest to lowest; however, as with all ordinal-level data, the exact difference in population between successive rankings would be unknown. On the other hand, a weakly ordered variable is based on nominal-level groups that are then rank-ordered in a meaningful arrangement. These groups represent frequency counts and do not display individual rankings of the aggregated data, thus limiting the amount of information conveyed when compared with strongly ordered variables. Categories showing country population as low (less than 1 million), medium (between 1 million and 100 million), and high (greater than 100 million), for instance, are weakly ordered because rankings within the ordered categories cannot be differentiated.

Numeric measurements on ordinal scales are ordered such that higher (lower) rankings are associated with larger (smaller) values. A common ordinal scale frequently employed in social and behavioral research is the Likert scale, which uses a hierarchical ranking system to indicate comparative levels of satisfaction, confidence, agreement, and so on, about a subject. Many mental constructs in psychology cannot be observed directly; therefore, these measures tend to be ordinal (e.g., Likert scaling). Opinions, attitudes, level of anxiety, specific personality characteristics, and so on, are all constructs that are regarded as varying in degree among individuals but tend to allow only indirect ordinal measurements. These are generally self-report measures. For example, a subject might be asked to rate the level of satisfaction he or she experiences with his or her current job on a Likert scale from 1 (*extremely dissatisfied*) to 5 (*extremely satisfied*). It cannot be assumed that a person assigning a rating of "4" to that question is exactly twice as satisfied with his or her job as a co-worker who answers the question with a "2." However, it is clear that the first person feels more satisfied with his or her job than the co-worker.

Providing rankings is also a common type of ordinal scale. For example, a subject might be asked to rank a list of values according to what is most important to him or her. Another example is a faculty search committee asked to rank-order a list of job candidates based on the overall qualifications in the areas of teaching ability, research skills, and so on. The candidate ranked "1" will be perceived by the committee as better than the candidate ranked "2," but not how much better. A special case of the ordinal scale occurs when data are used to classify individuals into two categories (a dichotomy), but the variable itself is assumed to be normally distributed with underlying continuity. For example, a psychological test might classify individuals into the categories *normal* (1) or *abnormal* (0). There is an implied order in the presence or absence of a characteristic.

As with other measurement scales, statistical methods devised to explicitly use ordinal-level data differ from approaches employing data in the nominal, interval, or ratio scales. Descriptive statistics, such as the median, range, and mode, can be determined for ordinal-level data in the same way as other measurement scales, but mean and standard deviation are not appropriate for ordinal data. This is because the numeric values are based on an arbitrarily defined starting point, and the distance between rankings is not identified. Correlation coefficients indicate the degree of relationship between one variable and another variable. The appropriate correlation coefficient for two ordinal-scale variables is the Spearman rank correlation coefficient or Spearman rho ($\rho$). Other correlations that include one variable at the ordinal level and the other variable at another level of measurement include rank-biserial (nominal and ordinal) and biserial (ordinal and interval/ratio). Several appropriate tests of significance are available for the one-sample case for ordinal data, including the Mann–Kendall tests for trends and the Kolmogorov–Smirnov one-sample test. Several nonparametric tests of significance are available for the two-sample case of ordinal data with independent samples, including the median test and the Mann–Whitney *U* test. In a multiple sample case with ordinal data, the Kruskal–Wallis one-way analysis of variance test is used. For example, a researcher might want to know if actors, musicians, and painters differ on a ranking of extroversion. The null hypothesis tested is that the population distributions from which the samples were selected are the same. If samples are dependent for ordinal data, the appropriate statistical

test of the two-sample case would be the Wilcoxon matched-pairs signed-rank test.

*Karen D. Multon and Jill S. M. Coleman*

***See also*** Interval Scale; Levels of Measurement; Likert Scaling; Mann–Whitney *U* Test; Nominal Scale; Nonparametric Statistics; Ratio Scale; Spearman Rank Order Correlation; Variable; Wilcoxon Rank Sum Test

**Further Readings**

Gravetter, F. J., & Wallnau, L. B. (2006). *Statistics for the behavioral sciences* (7th ed.). Belmont, CA: Wadsworth.

McGrew, J. C., Jr., & Monroe, C. B. (1999). *An introduction to statistical problem solving in geography* (2nd ed.). New York: McGraw-Hill.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103,* 677–680.

# Orthogonal Comparisons

The use of orthogonal comparisons within an analysis of variance is a common method of multiple comparisons. The general field of multiple comparisons considers how to analyze a multitreatment experiment to answer the specific questions of interest to the experimenter. Orthogonal comparisons are appropriate when the researcher has a clearly defined, independent set of research hypotheses that will constitute the only analyses done on the treatment means. Of the multiple comparison techniques available with comparable control over Type I error (rejecting a true null hypothesis), a set of orthogonal comparisons will provide a more powerful approach to testing differences that may exist between the treatment groups than those that are more exploratory in nature and thus more extensive. The power is gained by restricting the analysis to a small number of preplanned, independent questions rather than exploring for whatever difference may be present in the data.

Basic to an understanding of orthogonal comparisons is the concept of a comparison within a multigroup analysis of variance. Assume a balanced design in which each treatment group has an equal number of subjects. The great majority of comparisons deal with either differences between pairs of means, such as whether Group 1 differs from Group 2, or the differences between combinations of means, such as whether the average of Groups 1 and 2 differs from the average of Groups 3 and 4. This latter type of compound hypothesis, involving combinations of groups, is common in orthogonal comparisons. It is appropriate when groups have common elements that make the combination a meaningful grouping. If a learning experiment involves treatment groups that differ with respect to the type of prompts provided to students, there may be some groups that have in common that the prompts are delivered verbally, although they differ in other respects. Other treatment groups may have the prompts delivered in written form, again differing in other respects. A compound question focuses on the overall difference between the verbally delivered prompts and those where the prompts were delivered in writing.

The comparison is conceptualized as a set of weights, one for each of the $k$ treatment groups, $a_j = \{a_{j,1}, a_{j,2}, \ldots, a_{j,k}\}$, whose sum must be zero and whose pattern reflects the question being addressed. For example, the set $\{1 - 1\ 0\ 0\}$ in a four-group experiment is a comparison because $\sum a = 0$. It would be used to compare the first group with the second group, ignoring the third and fourth groups. A different comparison is defined by the weights $\{1,\ 1,\ -1,\ -1\}$, which also sum to zero and combine the first two groups and compare that sum with the sum of the last two groups. The important information in a set of weights is the pattern. The set $\{1/2, 1/2, -1/2, -1/2\}$ has the same pattern and would produce equivalent results to the set $\{1, 1, -1, -1\}$.

The null hypothesis for a comparison asserts that the sum of the products of weights multiplied by the population means is zero. Call this sum of products $\psi$. The null hypothesis is that

$$\psi = a_{11}\mu_1 + a_{12}\mu_2 + \cdots + a_{1k}\mu_k = 0 \qquad (1)$$

or, for the example above with four treatment groups,

$$\psi_1 = (1)\mu_1 + (1)\mu_2(-1)\mu_3 + (-1)\mu_4 = 0. \qquad (2)$$

If the null hypothesis is true and the overall effect of the first two groups is the same as that of the last two, the sample-based estimate of,

$$\hat{\psi}_1 = (1)\overline{X}_1 + (1)\overline{X}_2 + (-1)\overline{X}_3 + (-1)\overline{X}_4 \quad (3)$$

will randomly differ from zero. Tests of significance based on the probability of $\hat{\psi}_1$, given the null hypothesis that $\psi = 0$, can be evaluated using either the $t$ test or $F$ test with identical outcomes.

The concept of orthogonality is relevant when an experiment has more than one comparison. Experiments with multiple groups typically have a number of questions that are of interest. Groups that are combined in one comparison may be compared or combined in a different way in another. The particular set of comparisons chosen must reflect the specific questions of interest to the experimenter.

Imagine an experiment in which two specific comparisons are of interest. An issue of importance concerns the relationship between the two comparisons. The major distinction is between pairs of comparisons that are orthogonal and those that are not. The condition of orthogonality refers to the degree of independence of the questions. Comparisons that are orthogonal are independent and thus uncorrelated. They account for separate aspects of the total amount of variability between the means. The indication that two comparisons are orthogonal is that the sum of the products of the corresponding weights is zero (e.g., $\sum a_{1,j}\, a_{2,j} = 0$). The weight for Group 1 from Comparison 1 is multiplied by the weight for that group from Comparison 2. In a like manner, the weights for Groups 2 through $k$ for Comparison 1 are multiplied by the corresponding weights for Comparison 2. These products are then summed. Only if the sum of these products is zero are the comparisons said to be orthogonal.

Assume that there is a comparison of interest defined by the weights $a_1 = \{1, 1, -1, -1\}$. For this particular comparison, a number of additional comparisons ($a_2$) could be investigated that would be orthogonal to it. One of these is the comparison defined by $a_{2a} = \{1, -1, 0, 0\}$, which would look at the difference between the first two groups, a difference obscured in the first comparison by summing Groups 1 and 2 together. The sum of the products of the weights for these comparison is $\sum a_{1,j}\, a_{2a,j} = (1)(1) + (1)(-1) + (-1)(0) + (-1)(0) = 0$. Another comparison with weights of $a_{2b} = \{1, -1, 1, -1\}$ would also be orthogonal to $a_1$ as the sum of the products of weights is zero [e.g., $\sum a_{1,j}\, a_{2b,j} = (1)(1) - (1)(-1) + (-1)(1) + (-1)(-1) = 0$]. Additional comparisons would form an "orthogonal set" if each comparison were mutually orthogonal to each of the other comparisons in the set. There is no single test for a set of mutually orthogonal comparisons. Rather, an orthogonal set is one in which the sum of the products for each pair of comparisons is zero.

Comparisons that are nonorthogonal share common variability in that they share a portion of a common pattern. For example, $a_1 = \{1, 1, -1, -1\}$ and $a_2 c = \{1, 0, -1, 0\}$ both involve the difference between the first and third groups, in one case including Groups 2 and 4, and in the other, ignoring these groups. The sum of products is $\sum a_{1,j}\, a_{2c,j} = (1)(1) + (1)(0) + (-1)(-1) + (-1)(0) = 2$. Because the sum is not zero, the comparisons are not orthogonal.

Because of the mutual independence of the comparisons within a set of orthogonal comparisons, the possible number of comparisons contained within any specific set is limited to $k - 1$, the number of independent pieces of information or degrees of freedom associated with the sum of squares between groups. The sum of squares associated with a comparison is a part of the sum of squares between groups. In a complete set of $k - 1$ orthogonal comparisons, the sum of squares between groups will be partitioned into $k - 1$ comparison sums of squares, each with one degree of freedom, that add up to the sum of squares between groups.

In the example above, both $a_{2a} = \{1, -1, 0, 0\}$ and $a_{2b} = \{1, -1, 1, -1\}$ were orthogonal to $a_1 = \{1, 1, -1, -1\}$, but they are not orthogonal to one another (e.g., $\sum a_{2a,j} a_{2b,j} = 2$), so they do not form an orthogonal set. Because our example has $k = 4$ treatment groups, a complete set would have $k - 1 = 3$ comparisons. The partial sets defined by $a_1$ and $a_{2a}$ and by $a_1$ and $a_{2b}$ would have different third comparisons that uniquely complete the set. In the first case, the weights would be $a_{3a} = \{0, 0, 1, -1\}$, which would complete a hierarchical set, and in the second partial set, the weights would be $a_{3b} = \{1, -1, -1, 1\}$, which is part of a factorial set (see below).

Each comparison has a separate test of significance. Experimentwise (familywise) $\alpha$ Type I error control over the $k - 1$ comparisons in a complete set of orthogonal comparisons can be maintained by using a modification of Bonferroni's inequality by Zbynek Sidak. Per comparison control is sometimes reported, in which each comparison is evaluated at an $\alpha$ Type I error rate. The justification for this more liberal control rests on the small number of comparisons conducted, their independences, and their being planned prior to the conduct of the experiment.

Although the specific choice of orthogonal comparisons must reflect the interests of the researcher, there are several standard patterns. The following illustrations are based on $k = 4$ treatment groups. In a factorial design, if all treatment factors have 2 levels, the main effects and interaction(s) are a set of orthogonal comparisons. In a $2 \times 2$ with treatment groups $A_1B_1$, $A_1B_2$, $A_2B_1$, $A_2B_2$, the $A$ main effect, $B$ main effect, and $A \times B$ interaction would be given by the weights {1, 1, −1, −1}, {1, −1, 1, −1}, and {1, −1, −1, 1}, respectively. A hierarchical set consists of a compound comparison involving two combinations of groups. The remaining comparisons explore differences within the combinations. An example set of weights would be {1, 1, −1, −1}, {1, −1, 0, 0}, and {0, 0, 1, −1}. A Helmert set consists of a series of comparisons in which a single treatment group is compared with the remaining groups combined. The basic strategy of comparing one group to a combination of groups continues with the previous single group omitted. For four treatments, the weights would be {3, −1, −1, −1}, {0, 2, −1, −1}, and {0, 0, 1, −1}. The last common type is a set of orthogonal polynomials. The comparisons test for linear, quadratic, and higher order trends in the treatment means where the highest order in a particular set is $k - 1$. This set is specifically designed for treatments where the groups differ by amount rather than kind of treatment. With four treatments, it would be possible to test for linear, quadratic, and cubic trends. The weights would be {3, 1, −1, −3}, {1, −1, −1, 1}, and {1, −3, 3, −1}. Note that in all of these sets, there are $k - 1$ or three comparisons as there are $k = 4$ treatments.

A complete set of orthogonal comparisons provides an elegance of design in which all of the variability in the means is accounted for somewhere within the set of only $k - 1$ comparisons. However, there are occasions in which one or more additional comparisons may be needed to provide a satisfactory analysis of the data. This expanded set can be accommodated by using the critical value from Sidak that reflects the increased number of comparisons actually tested.

Orthogonal comparisons impose strict conditions on the number and nature of the comparisons within an experiment. For many experiments, orthogonal comparisons are too restrictive and do not allow the researcher to explore the many possible ways in which groups might differ. However, when the experimenter is willing to limit the questions asked to an orthogonal set of only $k - 1$ comparisons, it will translate into increased power and a more parsimonious description of the differences present.

*Alan J. Klockars*

***See also*** Analysis of Variance (ANOVA); A Priori Monte Carlo Simulation; Bonferroni Procedure; Multiple Comparison Tests; Pairwise Comparisons; Post Hoc Analysis; Type I Error

**Further Readings**

Hsu, J. C. (1996). *Multiple comparisons: Theory and method*. Boca Raton, FL: CRC.

Klockars, A. J., & Hancock, G. R. (1992). Power of recent multiple comparison procedures as applied to a complete set of planned orthogonal contrasts. *Psychological Bulletin, 111,* 505–510.

Klockars, A. J., & Sax, G. (1986). *Multiple comparisons* (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–061). Beverly Hills, CA: Sage.

Lomax, R. G. (2007). *Statistical concepts: A second course* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Maxwell, S. E., & Delaney, H. D. (2003). *Designing experiments and analyzing data* (2nd ed.). Belmont, CA: Wadsworth.

Sidak, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association, 62,* 626–633.

# OUTLIER

An outlier is an observation in a set of data that is inconsistent with the majority of the data. An

observation (i.e., score) is typically labeled an outlier if it is substantially higher or lower than most of the observations. Because, among other things, the presence of one or more outliers can dramatically alter the values of both the mean and variance of a distribution, it behooves a researcher to determine, if possible, what accounts for their presence. An outlier can represent a valid score of a subject who just happens to represent an extreme case of the variable under study, it can result from failure on the part of a subject to cooperate or follow instructions, or it can be due to recording or methodological error on the part of the experimenter. If the researcher has reason to believe that an outlier is due to subject or experimenter error, he or she is justified in removing the observation from the data. On the other hand, if there is no indication of subject or experimental error, the researcher must decide which of the following is the most appropriate course of action to take:

a. Delete the observation from the data. One disadvantage associated with removing one or more outliers from a set of data is reduction of sample size, which in turn reduces the power of any statistical test conducted although in the case of commonly employed inferential parametric tests (e.g., *t* test and analysis of variance), the latter may be counteracted by a decrease in error variability.

b. Retain the observation, and by doing so risk obtaining values for one or more sample statistics that are not accurate representations of the population parameters they are employed to estimate.

c. Retain one or more observations that are identified as outliers and evaluate the data through the use of one of a number of statistical procedures that have been developed to deal with outliers, some of which are discussed in the last part of this entry.

## Strategies for Identifying Outliers

One strategy for identifying outliers employs a standard deviation criterion that is based on Chebyshev's inequality, which states that regardless of the distribution of a set of data, $[1 - (1/k^2)](100)$ percent of the observations will fall within $k$ standard deviations of the mean (where $k$ is any value greater than 1). Because the percent values computed are quite high when $k$ is equal to 2, 3, 4, and 5 standard deviation units (respectively, 75%, 88.89%, 93.75%, and 96%), it indicates that scores that are beyond two standard deviations from the mean are relatively uncommon, and scores that are three standard deviations and beyond from the mean are relatively rare. However, employing a standard deviation criterion (e.g., declaring any observation that is three or more standard deviations from the mean an outlier) can be misleading sometimes, especially with small sample sizes and data sets with anomalous configurations.

Another tool that can be employed for identifying outliers is a box-and-whisker plot (also referred to as a *boxplot*), which is a visual method for displaying data developed by the statistician John Tukey. Typically, in a box-and-whisker plot, any value that falls more than one and one-half *hinge-spreads* outside a *hinge* can be classified as an outlier, and any value that is more than three hinge-spreads as a *severe outlier.* The value of a hinge-spread is the difference between the lower and upper hinge of the distribution. In a large sample, the upper and lower hinges of a distribution will approximate the 25th and 75th percentiles.

A final strategy for identifying outliers is to conduct an inferential statistical test. More than 50 different tests have been developed, most of which assume that a set of data is derived from a normal distribution. Some tests identify only the presence of a single outlier, whereas others are capable of identifying multiple outliers. Some tests are specific with respect to identifying outliers in one or both tails of a distribution. It should be noted that it is not uncommon for two or more tests applied to the same set of data to yield inconsistent results with regard to whether a specific observation qualifies as an outlier.

## Strategies for Dealing With Outliers

As noted earlier, aside from deleting observations that are identified as outliers from a set of data, a number of other strategies are available. A strategy referred to as *accommodation* uses all of the data, yet minimizes the influence of outliers. The latter can involve use of a nonparametric test that rank-orders data and uses the median in lieu of the

mean as a measure of central tendency. Another strategy is to evaluate the data with a *robust statistical procedure,* which is any one of more recently developed methodologies that are not influenced (or only minimally affected) by the presence of outliers.

Two other types of accommodation are *trimming* and *Winsorization.* Trimming involves deleting a fixed percentage of extreme scores from both the left and right tails of a distribution (e.g., the extreme 10% of the scores are removed from both tails). In Winsorization (named for the statistician Charles Winsor), a fixed number of extreme scores in a distribution are replaced with the score that is in closest proximity to them in each tail of the distribution. To illustrate, consider the following distribution of scores: 1, 24, 25, 26, 26, 28, 32, 33, 33, 36, 37, 39, 98. Winsorizing would be replacing the values 1 and 98 at the lower and upper end of the distribution with the values closest to them, 24 and 39.

*David J. Sheskin*

***See also*** Box-and-Whisker Plot; Range; Trimmed Mean; Winsorize

**Further Readings**

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* (2nd ed.). Chichester, UK: Wiley.
Rosner, B. (2006). *Fundamentals of biostatistics* (6th ed.). Belmont, CA: Thomson-Brooks/Cole.
Sheskin, D. J. (2007). *Handbook of parametric and nonparametric statistical procedures* (4th ed.). Boca Raton, FL: Chapman & Hall/CRC.
Sprent, P., & Smeeton, N. C. (2007). *Applied nonparametric statistical methods* (4th ed.). Boca Raton, FL: Chapman & Hall/CRC.
Tabachnick, B., & Fidell, L. (2007). *Using multivariate statistics* (5th ed.). Boston: Pearson/Allyn & Bacon.
Wilcox, R. (2003). *Applying contemporary statistical techniques.* San Diego, CA: Academic Press.

# OVERFITTING

Overfitting is a problem encountered in statistical modeling of data, where a model fits the data well because it has too many explanatory variables. Overfitting is undesirable because it produces arbitrary and spurious fits, and, even more importantly, because overfitted models do not generalize well to new data.

Overfitting is also commonly encountered in the field of machine learning, including learning by neural networks. In this context, a learner, such as a classifier or an estimator, is trained on an initial set of samples and later tested on a set of new samples. The learner is said to be overfitted if it is overly customized to the training samples and its performance varies substantially from one testing sample to the next.

Because the nature of the overfitting problem and the methodologies of addressing it are fundamentally similar in the two fields, this entry examines overfitting mostly from the viewpoint of statistical modeling.

## Relationship Between Model Fit and the Number of Explanatory Variables

In a statistical model, the fit of the model to the data refers to the extent to which the observed values of the response variable approximate the corresponding values estimated by the model. The fit is often measured using the coefficient of determination $R^2$, the value of which ranges between 0 (no fit) and 1 (perfect fit).

As one adds new explanatory variables (or *parameters*) to a given model, the fit of the model typically increases (or occasionally stays the same, but will never decrease). That is, the increase in fit does not depend on whether a given explanatory variable contributes significantly to the overall fit of the model or adds to its predictive power. Therefore, all other things being equal, a larger number of explanatory variables typically means better fit, but it does not necessarily mean a better model.

How many variables are too many depends on the model and the data. There is no preset number of variables above which the model is considered overfitted. Rather, overfitting is a graded property of a model: A given model is more or less overfitted depending on the number of parameters it has relative to the number of parameters actually needed.

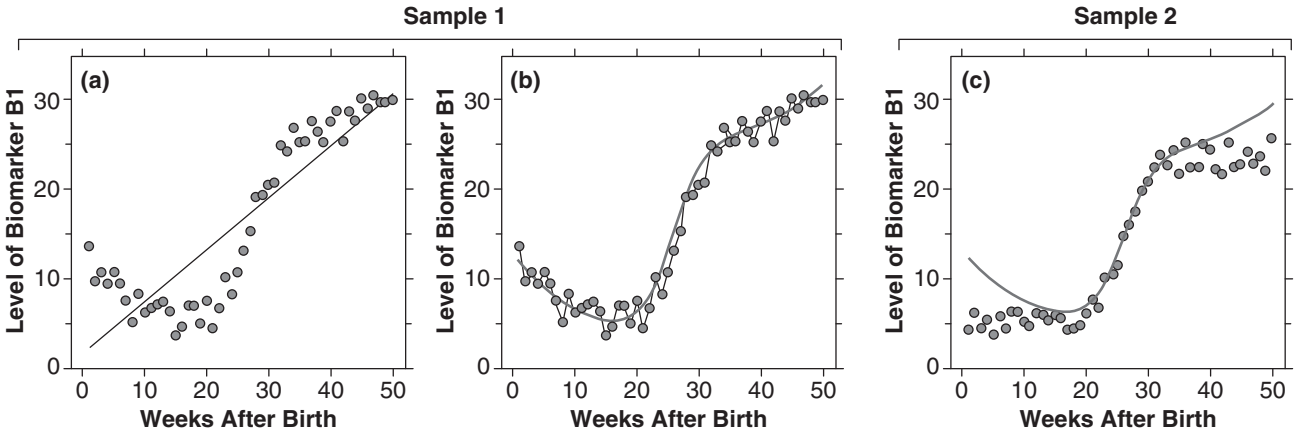



**Figure 1** Overfitting and Underfitting

*Notes:* These hypothetical clinical data show the blood levels of a biomarker B1 in a group of newborn infants measured during the first 50 weeks of their lives. Panels A and B are from the first group of babies (Sample 1), and the data in Panel C are from a different, independent sample. The dots represent the observed values of the biomarker (in arbitrary units). The thin black lines represent the underfitted and overfitted models (Panels A and B), respectively. The thick gray line in Panel B denotes the parsimonious model that best fits Sample 1. The same model is also shown in Panel C for comparison. Note that even though the model fits the data in Sample 1 rather well, it fails to predict the new data in Sample 2.

## Why Overfitting Is Undesirable

The function of a statistical model is twofold: First, it should help account for, or explain, the observed values in the original sample(s) on which it is based. Second, and more importantly, it should reliably predict the observed values in new testing samples. Overfitting compromises both the explanatory and predictive abilities of a model.

To understand how, suppose one is interested in modeling the levels of a hypothetical biomarker in healthy babies as a function of the baby's age (Figure 1). Age-appropriate levels of the biomarker signify normal development. The goal of the modeling endeavor is to determine which factors affect the levels of this biomarker.

Figures 1a and 1b show the biomarker levels measured in one group of babies during the first 50 weeks of their lives. When baby's age is used as the sole explanatory variable—that is, when the model underfits the data—the model is consistently "off the mark"; it generally misestimates the response variable (Figure 1a). In this case, the model is said to have high *bias*.

Figure 1b shows the same data overfitted by blindly using every last bit of information about the babies as explanatory variables—including age, gender, ethnicity, diet, and hours of sleep per day—up to and including clearly absurd variables such as the color of the diaper, number of traffic deaths during the given week, and the phase of the moon. This model, although manifestly absurd, fits the observed data from this sample nearly perfectly (Figure 1b). Yet the model explains little of the observed biomarker levels. That is, a good fit does not necessarily make a good model.

The model is whittled down to retain only those factors that can account for a statistically significant portion of the observed data. This more parsimonious model provides a reasonable approximation to the data (Figure 1b, thick gray line). Therefore, this model fulfils the first of its aforementioned functions: It helps explain the data at hand.

However, this model still does not provide a good fit to a new sample (Figure 1c). This is the second, larger problem with the overfitted models: Because they are overly customized to the original samples, they are not sensitive to data patterns that occur only in some of the samples if those patterns happen not to be present in the original samples. Therefore, they fail to generalize well across all samples, so that their degree of fit tends to vary from one sample to the next. Such models are said

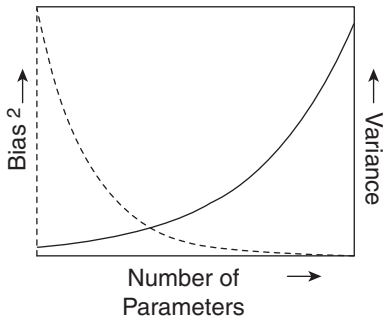



**Figure 2** Model Selection as a Trade-Off Between Bias and Variability

*Note:* As the number of model parameters increases, the bias of the model decreases (dashed line), and the variance of the model increases (solid line). The point at which the two lines intersect represents the most optimal theoretical balance between underfitting and overfitting (i.e., optimal bias-variance trade-off) when all other things are equal.

to have high *variance*. That is, the second model, although more parsimonious, is still overfitted.

## Preventing Overfitting: The Principle of Parsimony

Clearly, a good model avoids both of these extremes of underfitting and overfitting. The general approach to building models that avoid these extremes is enunciated by the *principle of parsimony,* which states that a model should have the fewest possible number of parameters for adequate representation of the data. Achieving parsimony by "shaving away all that is unnecessary" is often referred to as *Occam's Razor,* after the 14th-century English logician William of Occam.

### *Bias-Variance Trade-Off*

A principled, and widely used, approach to determining the optimal number of parameters in a model is to find the optimal trade-off point between bias and variance (Figure 2). As noted above, if the model has too few parameters (i.e., if it is underfitted), it tends to have a large bias. If it has too many parameters (i.e., if it is overfitted), it tends to have a large variance. Thus, the number of parameters that best balances bias with variance achieves the best balance between underfitting and overfitting.

It is important to note that the bias-variance trade-off represents only a conceptual guideline for preventing overfitting, and not an actual methodology for doing it. This is because, for one thing, it is impossible to precisely determine the point of optimal trade-off with a finite number of samples and without knowing the "ground truth," that is, what the "correct model" actually is. For another, the optimal balance between bias and variance often depends on many factors, including the goal of modeling. For instance, when simplicity of the model is paramount, a researcher may select the smallest number of parameters that still provide a significant fit to the data. On the other hand, when bias is more "costly" than variance, one may retain more parameters in the model than strictly necessary. Thus, a model that is not the most parsimonious is not necessarily a bad model.

### *Model Selection Methods*

Just as there is no single, universally applicable definition of what a good model is, there is no single, universally acceptable method of selecting model parameters. Rather, there is a variety of available methods. The methods vary widely in how they operate and how effective they are in preventing overfitting. It is beyond the purview of this entry to survey them all. Instead, some main approaches are briefly outlined.

As noted above, one can start with an overfitted model and test each explanatory variable for whether it contributes significantly to the model fit at a given alpha level (e.g., 0.05), and retain it in the model only if it does. Alternatively, one can start with an underfitted model, and incorporate only those variables that significantly improve the overall fit of the model. These step-up (or forward) versus step-down (or backward) parameter selection methods generally (although not always) produce largely similar models. Such significance-testing (or hypothesis-testing) methods, although widely used in the biological and social sciences, are considered inadequate by themselves for model selection. This is ultimately because these methods only address the ability of a given variable to explain the original data, and do not test the ability of the model to predict new ones. *Cross-validation* addresses some of these shortcomings, where the selected model is tested against new samples.

Akaike's Information Criterion (AIC) is an information theory-based index that measures the goodness of fit of a given model. It, along with its many variations, is a measure of the trade-off between bias and variance. For this and many other reasons, this measure has been a highly regarded method of preventing overfitting. Mallows' $C_p$ statistic is another goodness-of-fit measure that generally, but not always, produces results similar to AIC.

Using Bayesian principles of inference is one of the most promising new approaches to the prevention of overfitting. In machine learning, Bayesian principles have been used in two main ways to prevent overfitting. First, prior marginal probability distribution of a given explanatory variable has been used to help infer the extent to which the variable "belongs" in the model. Second, model selection itself can be treated as a Bayesian decision problem, so as to select a model that is most likely to be the "correct" model. These techniques, however, are yet to be widely used for preventing overfitting in the biological and social sciences.

It should be clear from the foregoing discussion that preventing overfitting is as much an art as it is a science because it involves exercising subjective judgment based on objective statistical principles. The fact that the optimal model is often in the eye of the beholder means that model selection methods can be abused to select one's favorite model. But such a fishing expedition of methods to support one's favorite model is inappropriate. The researcher's question ought to be which model best represents the data at hand, and *not* which method supports the model of choice.

*Jay Hegdé*

***See also*** Inference: Deductive and Inductive; Loglinear Models; Mixed Model Design; Models; Polynomials; SYSTAT

**Further Readings**

Akaike, H. (1983). Information measures and model selection. *Bulletin of the International Statistical Institute, 44,* 277–291.

Bishop, C. M. (2007). *Pattern recognition and machine learning.* New York: Springer.

Burnham, K. P., & Anderson, D. R. (2003). *Model selection and multi-model inference: A practical information-theoretic approach.* New York: Springer.

Draper, N. R., & Smith, H. (1998). *Applied regression analysis.* New York: Wiley.

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern classification.* New York: Wiley Interscience.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation, 4,* 1–58.

Harrell, F. E., Jr. (2001). *Regression modeling strategies.* New York: Springer.

Hawkins, D. M. (2004). The problem of overfitting. *Journal of Chemical Information and Computer Sciences, 44,* 1–12.

# P

## Pairwise Comparisons

Pairwise comparisons are methods for analyzing multiple population means in pairs to determine whether they are significantly different from one another. This entry explores the concept of pairwise comparisons, various approaches, and key considerations when performing such comparisons.

### Concept

Because population parameters (e.g., population mean) are unknown, practitioners collect samples for the purpose of making statistical inference regarding those parameters. As an example, many different statistical methods have been developed for determining if there exists a difference between population means. Perhaps most well-known is Student's $t$ test, which is typically used for inferring from two samples if there exists a difference between the corresponding population means (or, in the case of only one sample, determining if the population mean differs from some fixed constant). As an example, the researcher might be interested in performing a hypothesis test to determine if a particular reading improvement program among children is more effective than the traditional approach in a particular school district. The researcher might draw a sample of children to act as the control group and a similar sample to act as the treatment group. Gauging the effectiveness of the intervention by some measurement tool (e.g., words read per minute or reading comprehension), the researcher might implement a $t$ test to determine if there exists a significant difference between the control and the treatment group means.

In the case, however, that there are more than two means to be compared, the $t$ test becomes less useful. Consider a similar situation in which the researcher wishes to determine if there exists a difference between the effectiveness of several intervention programs. For example, the researcher might want to compare the effectiveness of several different programs that are implanted in each of, say, five school districts in a particular city. The researcher could, of course, simply perform multiple $t$ tests, comparing each of the intervention programs with all of the others. There are at least two critical problems with this approach.

The first is that the number of $t$ tests that are required to be performed dramatically increases as the number of treatment groups increases. In the case of five treatment levels and a single factor, there are only $\binom{5}{2} = 10$ different $t$ tests that would need to be performed. In the case, though, that there are multiple factors with multiple treatment levels, this number quickly becomes large. With the increase in computing power, however, this obstacle is not as meaningful as it has been in the past.

The second difficulty is more substantive. The second difficulty with performing multiple $t$ tests to compare several means is that it greatly

increases the false alarm rate (i.e., the probability of making a Type I error, which is rejecting the null hypothesis when it is, in fact, true). In the above example, suppose the level of significance is $\alpha = .05$, then the probability of making a Type I error is .05, and the probability of not making a Type I error is .95. Performing 10 $t$ tests, however, each with $\alpha = .05$, causes the probability of committing at least one Type I error to increase to $1 - .95^{10} = 0.4013$. That is, simply using $t$ tests to determine if there exists a difference between any of five population means, each with significance .05, results in around a 40% chance of committing a Type I error.

To alleviate this difficulty, practitioners use global tests of hypotheses. Techniques like analysis of variance (ANOVA) and multiple regression, for example, rely on the $F$ test, a global test, which tests the null hypothesis that all of the group means are equal against the alternative hypothesis that at least one of the group means is different from the others. By specifying the hypotheses in this way, the overall or experimental Type I error rate can be maintained because only one test is actually performed. The drawback is that the alternative hypothesis is much more vague. If the global test leads the researcher to reject the null hypothesis, the conclusion drawn is that the means are not all equal. The test gives no information as to which of the means might be different from which of the others. For example, all of the group means might be different from one another, or all of the group means might be equal except for one. Rejection of the null hypothesis in a global test of hypotheses might take place in either of these circumstances and would not allow the researcher to determine which is the case.

Once the determination has been made that the group means are not all equal (i.e., the null hypothesis in the global test has been rejected), the practitioner may use pairwise comparisons to determine which of the means are significantly different from one another. Stated in terms of hypotheses, if there are $k$ comparisons to be done, then there are $k$ sets of hypotheses of the form $H_0 : \mu_i = \mu_j$ and $H_1 : \mu_i \neq \mu_j$ for all $i \neq j, 1 \leq i, j \leq a$ where is the number of treatment groups in the study.

From a notational perspective, suppose that some experiment resulted in four sample means that, when sorted from least to greatest, are denoted by $\overline{x}_1$, $\overline{x}_2$, $\overline{x}_3$, and $\overline{x}_4$ and with corresponding population means denoted $\mu_1$, $\mu_2$, $\mu_3$, and $\mu_4$, respectively. Suppose further that some pairwise comparison procedure suggests that $\mu_1$ and $\mu_2$ are not statistically different from one another, but that they are statistically different from $\mu_3$ and $\mu_4$, which, in turn, are not found to be statistically different from each other. Typically, this is denoted as $\underline{\mu_1\ \mu_2}\ \underline{\mu_3\ \mu_4}$, where the means that those which share an underline are not found to be statistically different from one another by the pairwise comparison procedure.

At the heart of each of these procedures is the desire to avoid the problem of increased experiment-wide Type I error rate described above when performing multiple tests. That is, in determining which of the group means are different from one another, the practitioner would like to be able to maintain an overall, experiment-wide Type I error rate of $\alpha$; simply performing multiple $t$ tests of size $\alpha$ maintains only a per-statement or pairwise error rate of $\alpha$, and the overall error rate increases as the number of tests performed increases.

## Approaches

One of the very common approaches to pairwise comparisons is to apply a Bonferroni adjustment to the individual comparisons in order to preserve the experiment-wide Type I error rate. The probability of making a Type I error in the $i$th comparison is denoted as $\alpha_i$ and the experiment-wide error rate as simply $\alpha$. The Bonferroni inequality (also known as Boole's inequality) states that

$$\alpha \leq \sum_{i=1}^{k} \alpha_i,$$

where $k$ is the number of pairwise tests to be performed. If the practitioner (very reasonably) uses the same significance level for each pairwise comparison, then we have $\alpha \leq k\alpha_i$. In other words, the practitioner can guarantee that the overall Type I error rate is less than or equal to $\alpha$ by setting the pairwise significance level to $\alpha/k$, where equality is achieved in the case of independence. Hence, by reducing the significance level of each of the pairwise comparisons, the researcher can control the

experiment-wide error rate. However, the Bonferroni adjustment has been shown to be a very conservative approach that fails to allow the researcher to detect important differences between the means (i.e., the power is diminished by an overly conservative approach).

Fisher's least significant difference (LSD) procedure for pairwise comparisons was proposed by Ronald A. Fisher in 1935 and is among the oldest modern methods for pairwise comparisons. Fisher, who was the first to define and use the term *variance* and who is widely credited with having invented ANOVA, devised the LSD test to solve this problem as it relates to analysis of variance. Fisher's test uses a two-stage approach. First, the global test of hypothesis is considered. If there is not enough evidence to suggest that any of the treatment means are different in the global test, then the decision in the pairwise tests is made in favor of the null hypothesis. If the global test leads to rejection of the null hypothesis, then the practitioner essentially performs $k$ $t$ tests of size $\alpha$ to determine which means are different.

The reader may object that because no correction has been made to the significance level of the individual tests, an increasing experiment-wide Type I error rate will be encountered. Such an observation is correct. Fisher's LSD test does not control the experiment-wide Type I error rate. However, because multiple comparisons are performed only after rejecting the null hypothesis in the first stage, the increase in the Type I error rate is not as pronounced. Despite not controlling the experiment-wide Type I error rate, its simple and intuitive approach has helped it endure.

Tukey's test, also known as the Studentized range test or the honestly significant difference (HSD) test, does control the experiment-wide error rate. This procedure was originally developed for pairwise comparisons only when sample sizes among treatment groups are equal, but it was later adapted to handle treatment groups of unequal sizes. This version of the test, for the unbalanced case, is also known as the Tukey–Kramer procedure. In the balanced case, two population means are considered to be significantly different if the absolute value of the difference of the corresponding sample means is greater than

$$T_a = q_a(p,f)\sqrt{\frac{MSE}{n}},$$

where $p$ is the number of sample means in the study, $f$ is the degrees of freedom used in calculating mean square error (MSE), $n$ is the number of observations in each treatment group, and $q_a(p, f)$ is the $100(1 - \alpha)^{\text{th}}$ percentile of the distribution of the Studentized range statistic,

$$q = \frac{\overline{x}_{\max} - \overline{x}_{\min}}{\sqrt{MSE/n}},$$

where $\overline{x}_{\max}$ and $\overline{x}_{\min}$ refer to the largest and smallest sample means in the sample, respectively.

The naming of Tukey's HSD test communicates that Tukey's test is more conservative than Fisher's LSD test. This is because Tukey's test controls the overall error rate. For this reason, it is often preferred to Fisher's test.

These three procedures are, by far, the most common for performing general pairwise comparisons. A few other, more specialized techniques, however, bear mentioning. First, Dunnett's multiple comparisons with control (MCC) test, as its name suggests, is used when comparing means against a control. Similarly, Hsu's multiple comparisons with best (MCB) test is used for comparing the other groups with the best among those tested. Because both of these tests control the overall Type I error rate, by reducing the number of comparisons being performed, they increase the power of each of the pairwise tests.

Because the more comparisons made by the practitioner, the smaller the size of the individual tests must be in order to preserve the experiment-wide error rate, one approach is to reduce the number of tests to be made. Dunnett first proposed the test that bears his name, Dunnett's MCC test, in 1955 for the purpose of performing pairwise comparisons when the practitioner is interested only in comparing each treatment group with a control group. If, for example, the researcher were comparing various new training programs with what is currently being used, Dunnett's test would be useful in determining if any of those tests outperform the current standard. The procedure for Dunnett's test is a modification of the $t$ test and generally requires a computer to perform because the table of constants required is not readily

available, as it depends on the number of comparisons to be made, the sample sizes of treatment and control groups, and the experiment-wide error rate, and is therefore quite lengthy.

Hsu's MCB test, like Dunnett's test, reduces the number of tests being performed, but instead of comparing treatment groups against a control, groups are compared against the best (or worst in the current sample). As an example, suppose a health agency is attempting to determine which community health program is the most effective at preventing smoking among adolescents. The researcher would undoubtedly like to know the best of the available approaches. If it is the case, however, that the treatment group with the greatest effectiveness is not statistically different from the group with the second highest effectiveness (or even the third highest), then the planning and budgeting offices of the community health organization can understand their options better. Perhaps the test with the highest effectiveness has higher funding requirements, but the second best test is cheaper. A pairwise comparison procedure that can show that its effectiveness is not statistically different from that of the first would be of great interest. In such a case, Hsu's test is called for.

Hsu's test is largely based on Dunnett's MCC test. In fact, they both rely on the same constant and therefore use the same tables of values. The results are also interpreted in similar ways. Whereas Dunnett's MCC identifies significant differences between treatment groups and the control, Hsu's test identifies significant differences between treatment groups and the best, where the definition of best depends on the context of the problem. In the event that high values indicate what is best, the difference of interest is $\mu_i - \max_{i \neq j} \mu_j$. Conversely, where low values indicate what is best, the difference of interest is $\mu_i - \min_{i \neq j} \mu_j$, $i = 1 \ldots k$, where $k$ is the number of treatment effects. For simplicity, the example taken will be the former. In this case, a positive difference indicates that the $i$th mean is larger than all of the others, and a negative difference indicates that the $i$th mean is not larger than the others. This description is known as the unconstrained approach because the endpoints of the confidence intervals constructed in this approach are not artificially constrained by the practitioner.

Hsu's test, however, is often used to construct *constrained* confidence intervals in which the interval is forced to contain zero. Often, the practitioner is not interested in how much worse the nonoptimum solutions are from the best. Constrained confidence intervals may be used when this is the case. By forcing the interval to contain zero, this test achieves sharper inference.

Many more pairwise comparison procedures have been developed, but they are, in large part, merely variants of those that have been described above. One approach that differs from the abovementioned approaches is the Waller-Duncan Test. This test implements a Bayesian framework and determines pairwise differences by minimizing the Bayes risk rather than explicitly controlling the experiment-wide Type I error rate. It allows the practitioner to specify a constant, $k$. This constant is used to make it more (or less) difficult to decide in favor of (or against) the null hypothesis, depending on the potential cost of making the wrong decision (i.e., committing a Type I or Type II error).

## Key Considerations

Three key considerations are important when performing pairwise comparisons: First, pairwise comparisons should be performed only after the null hypothesis in the global test of hypotheses has been rejected. Failing to reject this global hypothesis should also be considered a failure to find any of the pairwise comparisons significant. Performing repeated tests in a (frantic) search for significance will, no doubt, inflate the probability of making a Type I error. Second, care should be paid to the experiment-wide Type I error rate because of the effect of performing multiple tests. Even when using Fisher's test, which does not control the experiment-wide Type I error rate, the practitioner can apply a Bonferroni adjustment to do so. Finally, the quantity of tests can be reduced (thereby increasing the power of the multiple comparisons) by making only comparisons of interest. This is typically done using tests like Dunnett's MCC or Hsu's MCB test, but can also be accomplished simply by making fewer comparisons with Fisher's LSD test and adjusting $\alpha$ accordingly.

*Gregory Vincent Michaelson and Michael Hardin*

*See also* Mean Comparisons; Multiple Comparison Tests; Student's *t* Test; *t* Test, Independent Samples; *t* Test, Paired Samples; Type I Error

### Further Readings

Carmer, S. G., & Swanson, M. R. (1973). *Journal of the American Statistical Association, 68*(341), 66–74.
Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh, 52,* 399–433.
Fisher, R. A. (1921). Studies in crop variation. I. An examination of the yield of dressed grain from Broadbalk. *Journal of Agricultural Science, 11,* 107–135.
Hsu, J. C. (1996). *Multiple comparisons: Theory and methods*. London: Chapman & Hall.
Miller, R. G. (1991). *Simultaneous statistical inference* (2nd ed.). New York: Springer-Verlag.
Montgomery, D. C. (2004). *Design and analysis of experiments* (6th ed.). New York: Wiley.
Waller, R. A., & Duncan, D. B. (1969). A Bayes rule for the symmetric multiple comparisons procedure. *Journal of the American Statistical Association, 64,* 1484–1503.

## Panel Design

A panel design is used when researchers sample a group, or panel, of participants and then measure some variable or variables of interest at more than one point in time from this sample. Ordinarily, the same people who are measured at Time 1 are measured at Time 2, and so on. The successive measures are commonly referred to as *waves*. For example, a three-wave panel study would measure the same sample of participants on three separate occasions. The amount of time in between measurements is known as the *interwave interval*. The use of multiple measures on the same variable(s) over time allows for an assessment of longitudinal changes or stability on the variables of interest.

Perhaps the simplest and most common version of the panel design is the two-wave, two-variable panel, or *cross-lagged panel design*. In this design, two variables ($X$ and $Y$) are measured at two discrete points in time from the same sample of participants. The procedure allows the researcher to test a series of different associations depicted in Figure 1.

The horizontal paths in the cross-lagged panel model represent *stability coefficients*. These values indicate how stable the $X$ or $Y$ variables are over time. If analyzed with regression, the larger these regression coefficients are, the greater the stability of the $X$ or $Y$ constructs over the course of the interwave interval. The diagonal paths are of particular interest as they are often compared to determine if $X_1 \rightarrow Y_2$ is greater or less than $Y_1 \rightarrow X_2$. This comparison often forms the basis for making causal inferences about whether $X$ causes $Y$, $Y$ causes $X$, or perhaps some combination of the two. Although cross-lagged panel designs are very common in the literature, there is a considerable controversy about their actual utility for testing causal inferences. This is because a valid causal analysis based on a cross-lagged panel design involves the absence of measurement error, correct specification of the causal lag between T1 and T2, and the assumption that there is no third variable, $Z$, that could exert a causal impact on $X$ and $Y$. Rarely, if ever, are all of these assumptions met.

Despite its limitations, the cross-lagged panel study illustrates several key benefits of panel designs more generally. First and foremost, one of

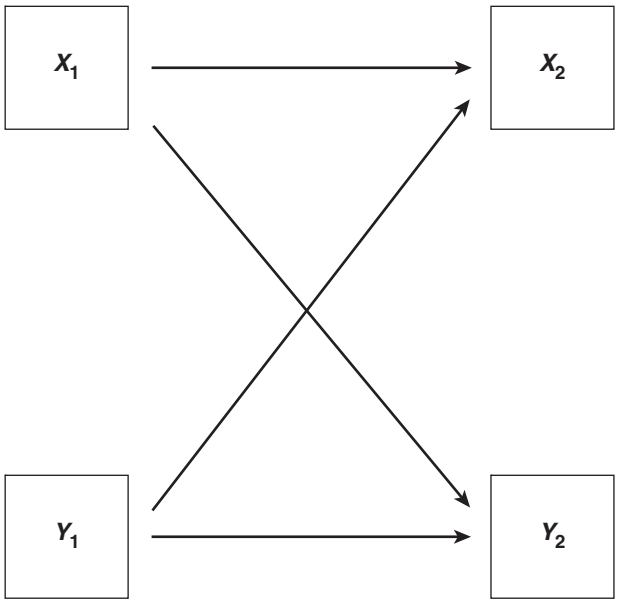



**Figure 1** Basic Structure of Cross-Lagged Panel Design

the fundamental prerequisites for establishing that *X* has a causal effect on *Y* is the demonstration that *X* occurs before *Y*. For example, if smoking is assumed to cause lung cancer, it is important to be able to demonstrate that people first smoke while otherwise free of cancer, and then lung cancer appears at a later point in time. Merely demonstrating that *X* precedes *Y* in a longitudinal panel study does not prove that *X* causes *Y,* but it satisfies this vital criterion, among several others, for establishing a causal relation. Another important quality of panel designs is that they do not confound interindividual growth with interindividual differences in intraindividual growth, the way that cross-sectional studies do. Because the same people are measured at multiple points in time, it is possible to statistically model changes within the same person over time, differences between people, and interindividual differences in the rate and direction of change over time.

The remainder of this entry addresses important design considerations for studies using panel designs, variants on the prospective panel design, and special problems associated with panel designs.

## Panel Study Design Considerations

There are several important research design considerations that must be considered when planning a study with a panel design. One potential problem with interpreting the results of a two-wave panel study is *regression toward the mean*. Over time, extreme scores on any variable have a natural tendency to become less extreme. This statistical artifact has the potential to obscure meaningful longitudinal associations, or lack thereof, between variables in the study. Concerns about regression toward the mean can be alleviated if more than two waves of measurement are taken. For this reason, it is generally advisable to implement three or more waves of measurement in any panel study.

Another important design consideration is proper specification of the *causal lag*. Quite simply, this is a question of how long it takes the effect of *X* to appear on *Y*. In other words, how fast will *Y* change as a function of *X*? The answer to this question has major implications for selecting the appropriate interwave interval. Misspecification of this interval can invalidate the results of a panel study. Assume, for example, that a researcher is interested in testing the effectiveness of a new exercise program on people's weight. An exercise physiologist determines that the initial effect of the program should become evident in approximately 2–3 weeks. If a researcher designs a test of this program's effectiveness with a three-wave panel study that has a 5-day interwave interval, it is likely that the effectiveness of the program will be missed all together. This is because the overall window of observation is too short to allow for the effect of the exercise program to show up on participants' weight. An interwave interval that is too long can be equally deleterious to testing causal inferences. Assume that a researcher wants to study the effects of losing a family member on people's level of psychological distress. In order to allow for a reasonable probability of observing the event of interest, a 15-year panel study is designed with measures every five years. It is entirely possible that psychological distress might shoot up shortly after the death of a family member, but the designer of this hypothetical study would miss that entirely because the person's psychological distress might return back down to normal levels by the time of the next wave of measurement, because the observations were spaced too far apart. For these reasons, it is vital that those who conduct panel studies have a clear understanding about the causal lag, or time ordering, between the variables of interest and use appropriate interwave intervals that allow for detection of these causal effects. As a general rule, the time lag between cause and effect should be shorter than the interwave interval. Ideally, the timing of observations will be dictated more by a theory of temporal change on the variable(s) of interest rather than logistics.

Related to specification of the causal lag is selection of the *number of waves*. Because of regression toward the mean, two-wave designs are generally less than ideal. However, it is not always clear whether there should be 3, 4, 5, or even 20 waves of measurement. Ordinarily, a higher number of waves is indicated when (a) the causal lag is relatively brief and (b) the probability of the causal event (*X*) occurring or causal variable (*X*) fluctuating that will ultimately influence the dependent variable (*Y*) is relatively low. The use of a larger number of waves will increase the

likelihood that the researcher will capture or observe this causal process while the study is ongoing.

### Variants on the Prospective Panel Design

An alternative to the prospective panel design is the *retrospective panel design*. In the retrospective panel design, data are collected retrospectively, with respondents providing information on a state of affairs at a point in time prior to the actual data collection. One attractive element of retrospective panel designs is the ability to collect what is essentially longitudinal data at a single point in time. For example, a researcher might be interested in how seat belt usage has changed as a function of laws and manufacturing trends in the auto industry. Perhaps the researcher wants to study seat belt use in 1975 prior to seatbelt legislation; in 1990, when automatic seatbelts started to become standard equipment in many cars; and in 1995, when airbags started to appear in most vehicles. One way to study this would be to obtain a sample of older adults and ask them to estimate what percent of the time they wore seatbelts during these various years. In effect, the researcher could obtain three waves of data, covering a 20-year period, from a single data collection session. In theory, there should be no attrition as all participants sampled for the study could provide data for all prior points in time.

However, this example also illustrates two important problems with retrospective panel designs. The first and most obvious problem stems from problems associated with recall. Memory for how often one wore seatbelts more than 30 years ago might be vague at best. Memory failure can produce both false positive (recalling use of seatbelts when the respondent did not actually use them) and false negative (failing to recall the use of seatbelts during a period in which they were actually used) recall errors. Another, less obvious problem with retrospective panel designs is censorship associated with survival. Assuming that those who never or rarely wear seatbelts are more likely to die in an auto accident, a sample of adults able to report on 30 years of seatbelt use will be biased toward those who are more likely to wear seatbelts. Those who used seatbelts less frequently are less likely to have survived over the interval of observations and are therefore less likely to appear in the study's sample. The retrospective panel design described in this example is a *follow-back* panel design whereby cases are sampled in the present and data from earlier points in time are generated by the participants. In the *catch-up* panel, a cross-sectional sample is secured on the basis of data from an archival source (e.g., men who registered for the selective service in 1980), and those cases are located at the present time to generate data for the present, and perhaps for the period in between the index event and the present.

Another variant of the prospective panel design is known as a *rotating panel design*. In this design, data are collected repeatedly from a sample of cases. However, over subsequent waves, some cases are dropped and replaced with new cases. One of the desirable qualities of the rotating panel design is minimization of problems associated with attrition, measurement, and fatigue. For example, assume that an economist is interested in studying the association between the market value of various precious metals and the welfare of residents in mining communities. Because of the volatile nature of the precious metals market, it is desirable to conduct assessments at least once every 6 months. However, over a period of many years, repeated measurements from the same participants could become a burden, plus some people are likely to drop out of the study. These effects can be counteracted by recruiting a new, but comparable, sample to replace those who are dropped from the study after some predetermined time frame (e.g., 2 years). So long as the cause $\longrightarrow$ effect time lag is shorter than the period for which the same set of participants remains in the study—in this case, 2 years—causal inferences can be evaluated with such data. This is because short-term change on the individual level can be evaluated with a subset of the data. At the same time, long-term change at the group level is possible due to the overall homogeneity of the sample (e.g., all residents of mining communities).

Closely related to the revolving panel design is the *accelerated longitudinal design* first described by Richard Bell in 1953. In the accelerated longitudinal design, multiple panels are recruited and followed over a specified time interval. These panels have the quality of being uniformly staggered on key developmental stages. For example, assume

that a linguist wants to study the depth and breadth of language acquisition between ages 3 and 15. However, the linguist does not have 12 years to conduct the investigation. In the accelerated longitudinal design, the linguist could recruit four panels of children who are 3, 5, 7, and 9 years old and follow them all for a period of 3 years. In so doing, the researcher could collect 12 years' worth of developmental data in a period of only 3 years. One might wonder, why not collect data on 3-, 6-, and 9-year-olds for 3 years? This is because ideally there should be some overlap between the cohorts on the developmental variable on which they are staggered, in this case, age. In this example, data on language acquisition during the fifth year of life would be available from both Panel 1 and Panel 2, albeit at different times during the course of the study. This ability to link the panels plays an important role in the causal inference process. A major advantage of accelerated longitudinal designs is the ability to study aging effects independent of period effects. Over a 12-year period, societal changes in instructional technology, computers, and so on might account for greater depth and breadth of language acquisition. By accelerating the collection of 12 years' worth of data to only 3 years, these period effects are minimized. A serious drawback of accelerated longitudinal designs is that intra-individual changes can be studied only over much shorter intervals than what is represented in the totality of the design. In cases where there are longer term causal effects, this can pose a very serious problem for testing causal inferences. Accelerated longitudinal designs also assume that a single growth trajectory is sufficient for describing all cohorts or panels, and this assumption might not be valid in some cases.

## Special Problems With Panel Designs

Notwithstanding their numerous benefits, panel studies have a series of special problems that must be taken into consideration. First, panel studies have unique problems associated with *attrition,* which is sometimes known as *mortality.* Attrition occurs when people who started the study at Time 1 eventually drop out before the final wave of measurement. Obviously, this is largely a problem with prospective panel designs, in contrast to retrospective panel designs. The longer the study and the more waves of assessment, the greater the likelihood of experiencing significant attrition. Attrition can introduce systematic bias into the data set generated by a panel study. For example, a researcher might want to test the effectiveness and perseverance of a drug rehabilitation program by following a sample of people who have recently completed the rehabilitation program and assessing their substance use and well-being every year for 10 years. At the end of the study, the rehabilitation program might appear to have been a resounding success. However, this could potentially be an artifact of attrition. What if the reality was that 35% of the cases relapsed and dropped out of the study due to poor health, hospitalization, incarceration, or even death from overdose? Only the well-functioning cases would be left in the study. This could easily lead the researcher to the wrong conclusion about the program because perseverance in the study is confounded with well-being.

In very long-term panel studies that sometimes take place over decades, the state of the art for *measurement* of the study's variables might change dramatically. Indeed, the very conceptual definition of the study's constructs (e.g., domestic violence, ideal body weight, mental illness) might change over time. The researcher is thus faced with the dilemma of using the old and outdated definition and measure that were used at Time 1, or to adopt the new, state-of-the-art measure at a later wave of assessment, therefore assuming the risk that the two different measures might not be comparable. In such cases, it might be unclear if a key variable changed meaningfully over time, or if the two different measures just produced different results.

*Testing effects* can occur any time participants are subjected to repeated assessments of the same or a similar set of variables over time. If an epidemiologist wanted to study the effects of smoking and drinking on the risk for developing certain health problems, participants in the study might adjust their levels of smoking and drinking merely because they know that they will be asked questions about these behaviors in the near future and are thus more consciously aware of them and perhaps their negative potential consequences.

In retrospective panel designs, *selection* can be a substantial problem. A retrospective study of

smoking and heart disease will left-censor cases of heavy smokers who had heart disease and did not survive up to the index date when the sample was selected. In effect, 100% of the sample will appear to have survived in the retrospective panel study, when in real life, this would not be the case. Even prospective panel studies might encounter problems with selection. Informed consent requires that participants are fully informed of the nature of an investigation prior to volunteering for it. Prospective panel studies that take place over very long periods of time with numerous waves of data collection might filter out certain people who do not want to assume that burden.

A final problem with any panel design aimed at testing causal inferences is the *proxy variable problem*. This happens when it is assumed from the analysis of the panel study data that *X* causes *Y*. However, in reality, *X* might merely be a proxy for some other variable, *W,* that is the true cause of *Y*. Longitudinal panel studies that tested the effects of parental divorce on children's well-being revealed that the occurrence of parental divorce is predictive of subsequent decreases, albeit small in magnitude, of children's psychological well-being. However, more careful analyses revealed that parental divorce was a proxy for such other variables as (a) parental conflict, (b) parental absence, and (c) lower family socioeconomic status, all of which can exert a causal effect on children's psychological well-being independent of parental divorce. The proxy variable problem is yet another illustration of why it is important to have a strong theory that specifies how, why, and under what conditions the cause-and-effect variables are linked.

*Chris Segrin*

***See also*** Cause and Effect; Cohort Design; Latent Growth Modeling; Longitudinal Design; Multilevel Modeling; Path Analysis; Sequential Design

**Further Readings**

Farrington, D. P. (1991). Longitudinal research strategies: Advantages, problems, and prospects. *Journal of the American Academy of Child and Adolescent Psychiatry, 30,* 369–374.

Kessler, R. C., & Greenberg, D. F. (1981). *Linear panel analysis: Models of quantitative change*. New York: Academic Press.

Menard, S. (1991). *Longitudinal research*. Newbury Park, CA: Sage.

Rogosa, D. (1979). Causal models in longitudinal research: Rationale, fomulation, and interpretation. In J. R. Nesselroade & P. B. Bates (Eds.), *Longitudinal research in the study of behavior and development* (pp. 263–302). New York: Academic Press.

# PARADIGM

Originally, *paradigm* was a word referring to an accepted model or pattern. In *The Structure of Scientific Revolutions,* Thomas Kuhn gave this word a new meaning, using it to represent the set of practices that constitutes the core of a scientific discipline and serves as a model to show how a scientific community should conduct its research.

In *Structure,* Kuhn defined a paradigm simply as a group of exemplary problem solutions universally accepted by the members of a scientific community. According to Kuhn, what constitute a paradigm are not abstract theories, but concrete applications of the theories for solutions of typical problems. For example, the paradigm of Newtonian mechanics does not consist of Newton's equations of motion, but exemplary applications of these equations in solving such standard problems as free fall, simple pendulum, and inclined planes. By forming analogies to these exemplary problem solutions, scientists conduct their research in a coherent way and establish a consensual core for the discipline. Initially, Kuhn did not think that paradigms exist in social sciences, because social scientists usually belong to competing schools and they are unable to reach universal consensus on what a common practice should be in their field.

Kuhn's notion of paradigm evolved after the publication of *Structure*. One development was that Kuhn gradually turned "paradigm" into a more inclusive concept. In the postscript of *Structure,* Kuhn introduced the notion of "disciplinary matrix" to replace the paradigm concept in the broad sense. The term *matrix* suggests that it is composed of various elements. In addition to

exemplary problem solutions, a disciplinary matrix also includes the following three components:

1. *Symbolic generalizations.* These are formal or formalizable expressions used by a scientific community as universal laws or equations, such as the symbolic form $f = ma$ defined by the second principle of Newtonian mechanics.
2. *Models.* These include heuristic analogies that help scientists to analyze and understand certain phenomena, such as using a hydrodynamic system as an analogy of an electric circuit. These also include metaphysical analogies used to justify certain ontological assumptions, such as using a group of tiny elastic billiard balls in random motion to illustrate the nature of gas.
3. *Values.* These are convictions shared by a scientific community, such as the belief in the significance of accuracy or simplicity.

Another development was that Kuhn eventually abandoned universal acceptance as the requirement for paradigms. Kuhn realized that scientific schools competing with each other may also have their own internal consensuses, and he agreed to label the consensus of a school a paradigm. After this revision, the concept of paradigm becomes applicable in social sciences.

## Paradigm Shifts

Kuhn called research conducted within the framework defined by a paradigm *normal science*. Routine research activities during normal science include gathering facts, matching facts with theories, and articulating theories. The aim of research in this stage is not to produce major novelties, but to solve puzzles. Scientists are confident that the paradigm can offer them all the necessary tools to solve puzzles, and they usually blame themselves rather than the paradigm when they encounter anomalies, that is, puzzles that have resisted solutions.

If some anomalies are persistent or challenge the fundamentals of the paradigm, they can eventually erode scientists' faith in the paradigm. The result is a crisis, which is frequently followed by a paradigm shift. A scientific revolution occurs, through which a new paradigm replaces the old one.

The most important contribution that Kuhn made in his study of paradigm shifts is his thesis of incommensurability. Kuhn believed that rival paradigms during a scientific revolution are incommensurable, in the sense that scientists from rival paradigms cannot communicate with each other without loss. In *Structure,* Kuhn uses Gestalt shifts as an analogy to illustrate his incommensurability thesis. According to Kuhn, after a revolution, those who belong to the new paradigm see things in an entirely different way, as if they were wearing glasses with inverting lenses. The Gestalt analogy implies a complete communication breakdown between scientists who belong to rival paradigms.

Kuhn later modified his position. In the postscript of *Structure,* he dropped the Gestalt analogy, abandoning the implied perceptual interpretation of the thesis. He instead developed a metaphor based on language: During scientific revolutions, scientists experience translation difficulties when they discuss terms from a different paradigm, as if they were dealing with a foreign language. Incommensurability is confined to meaning change of concepts, and becomes a sort of untranslatability.

In 1983, Kuhn introduced a notion of "local incommensurability" to narrow the scope affected by revolutions. During a scientific revolution, most of the terms function the same way in both paradigms and do not require translation. Meaning change occurs only to a small subgroup of terms, and problems of translatability or incommensurability are merely a local phenomenon.

Continuing this direction, Kuhn in 1991 further limited the scope of incommensurability by introducing a theory of kinds: Meaning change occurs only to a restricted class of kind terms or taxonomic terms. In this way, Kuhn redraws the picture of scientific revolutions. Because the interconnections among kind terms form a lexical taxonomy, scientific revolutions, which now are limited to the meaning change of kind terms, become taxonomic changes. A scientific revolution produces a new lexical taxonomy in which some kind terms refer to new referents that overlap with those denoted by some old kind terms. Scientists from rival paradigms face incommensurability because they construct different lexical taxonomies and thereby classify the world in different ways.

### The Progress of Science

One implication of the incommensurability thesis is that it is impossible to evaluate rival paradigms rationally, because there are no shared evaluation standards between those belonging to rival paradigms. In *Structure,* Kuhn affirmed this relativist position by citing the famous remark from Max Planck: "A new scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents eventually die, and a new generation grows up that is familiar with it" (p. 151).

Kuhn later tried to retreat from this relativist position by developing an analogy between scientific development and biological evolution. In *Structure,* Kuhn already used an analogy to biological evolution to illustrate scientific revolutions. Scientific development is parallel to biological evolution in the sense that both are products of competition and selection. Scientific development is a process driven from behind, not pulled from ahead to achieve a fixed goal. Kuhn further explicated this analogy in 1992, revealing another important similarity: Both scientific development and biological evolution have the same pattern of growth in the form of an evolutionary tree. Whereas proliferation of species is the form of biological evolution, proliferation of specialized disciplines is the key feature of scientific progress. The pattern of knowledge growth is an inexorable growth in the number of distinct disciplines or specialties over the course of human history.

With the analogy to biological evolution, Kuhn proposed an evolutionary epistemology to specify the evaluation standards for scientific development. The key of this new epistemology is to distinguish between evaluation of stable belief and evaluation of incremental change of belief. When we evaluate the development of two beliefs, it makes no sense to say that one is "truer" than the other, because there is no fixed platform that can supply a base to measure the distance between current beliefs and the true belief. To evaluate the incremental change of belief, one must apply a different criterion, the proliferation of distinct specialties. Thus, although incommensurability may continue to create inconsistent evaluation standards between rival communities, this confusion does not necessarily lead to a total failure of rational comparisons. The evolution of knowledge consists of the proliferation of specialties, the measurement of which is not only practicable but also frequently independent of the theoretical positions of evaluators. A rational evaluation of knowledge development is possible even when people employ different evaluation standards.

*Xiang Chen*

***See also*** *Logic of Scientific Discovery, The;* Positivism

**Further Readings**

Kuhn, T. (1962). *The structure of scientific revolutions.* Chicago: University of Chicago Press.

Kuhn, T. (1969). "Postscript—1969." In T. Kuhn, *The Structure of Scientific Revolutions* (2nd ed., pp. 174–210). Chicago: University of Chicago Press.

Kuhn, T. (1983). Commensurability, comparability, and communicability. In P. Asquith & T. Nickles (Eds.), *PSA 1982, Vol. 2* (pp. 669–688). East Lansing, MI: Philosophy of Science Association.

Kuhn, T. (1991). The road since *Structure.* In A. Fine, M. Forbes, & L. Wessels (Eds.), *PSA 1990, Vol. 2* (pp. 3–13). East Lansing, MI: Philosophy of Science Association.

Kuhn, T. (1992). *The trouble with the historical philosophy of science: Robert and Maurine Rothschild Distinguished Lecture.* An occasional publication of the Department of the History of Science, Harvard University.

## Parallel Forms Reliability

Parallel forms reliability (sometimes termed *alternate forms reliability*) is one of the three primary classifications of psychometric reliability, along with test–retest reliability and internal consistency reliability. The three categorizations of reliability differ primarily in the differing sources of non-trait or non-true score variability, and secondarily in the number of tests and occasions required for reliability estimation. This entry discusses the construction of parallel forms and the assessment, benefits, and costs of parallel forms reliability.

## Construction of Parallel Forms

The creation of parallel forms begins with the generation of a large pool of items representing a single content domain or universe. At minimum, the size of this item pool should be more than twice the desired or planned size of a single test form, but the item pool should also be large enough to establish that the content domain is well represented. Parallel test forms are generated by the selection of two item sets from the single universe or content domain. The nature of this selection procedure can vary considerably in specificity depending on setting, from random selection of items drawn from the homogeneous item pool, to paired selection of items matched on properties such as difficulty. The resulting forms will typically have similar overall means and variances.

A subtle distinction can be drawn between the concept of parallel forms as it is popularly used and the formal psychometric notion of parallel tests. In the context of classical test theory, parallel tests have identical latent true scores, independent errors, and identical error variances. In practice, two forms of a test considered to be parallel do not meet the formal standards of being parallel tests, but often approach such standards depending on procedures used in the item assignment process.

Parallel forms represent two distinct item sets drawn from a content domain. Multiple test forms that contain nondistinct sets (i.e., overlapping items) are useful in many situations. In educational settings, multiple test forms can be created by scrambling items, which ideally discourages cheating via the copying of nearby test answers. In this case, the two forms are actually the same form in terms of content, and the appropriate reliability assessment would be test–retest. Identical items may also be placed on two test forms to facilitate a comparison of their characteristics under differing conditions (e.g., near start vs. end of test).

## Assessment

Parallel forms reliability is assessed by sequentially administering both test forms to the same sample of respondents. The Pearson correlation between scores on the two test forms is the estimate of parallel forms reliability. Although the administration procedures for parallel forms reliability differ from test–retest reliability only in the existence of a second test form, this has important implications for the magnitude of the reliability coefficient. In the estimation of parallel forms reliability, error variance is composed not only of transient error occurring between Time 1 and Time 2, but also of variance due to differing item content across forms. Thus, test–retest reliability tends to be greater in magnitude than parallel forms reliability. A negligible difference between the two estimates would suggest minimal contribution of varying item content to error variance, and strong evidence supporting the use of the forms.

## Benefits and Costs

The development of parallel forms affords several advantages in both research and applied settings. Generally speaking, the existence of multiple forms sampled from a given content domain is useful whenever sampling or design issues preclude the use of a single test form. Populations differing in primary spoken language, or who have visual or hearing impairments, can be assessed using specially designed measurement instruments that will yield comparable scores with other populations. Practice effects, or other types of accumulated error that often confound situations where subject progress is being assessed, can be reduced when one or many parallel forms are administered at later time points in the assessment.

These potential advantages do not come without concomitant costs. Investments of time and money can effectively double when generating parallel forms, and the utility of these investments must be weighed against the potential gains from developing a measure of a different content area. In addition to costs involved in the development of the forms themselves, the content domain must be sampled more deeply, likely requiring a broader and more careful delineation of the domain boundaries, as well as increased care taken not to oversample subsets that may exist within the domain.

*William M. Rogers*

*See also* Correlation; Error; Internal Consistency Reliability; Test–Retest Reliability

**Further Readings**

Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). New York: HarperCollins.
Cronbach, L. J., & Azuma, H. (1962). Internal-consistency reliability fomulas applied to randomly sampled single-factor tests: An empirical comparison. *Educational and Psychological Measurement, 22,* 645–665.
Gibson, W., & Weiner, J. (1998). Generating random parallel test forms using CTT in a computer-based environment. *Journal of Educational Measurement, 35,* 297–310.
Lord, F. (1957). Do tests of the same length have the same standard errors of measurement? *Educational and Psychological Measurement, 17,* 510–521.
Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.
Sax, G., & Carr, A. (1962). An investigation of response sets on altered parallel forms. *Educational and Psychological Measurement, 22,* 371–376.
Thurstone, L. L. (1931). *The reliability and validity of tests.* Ann Arbor, MI: Edwards.

# PARAMETERS

Parameters are characteristics of populations—they describe the population. In research design, one is often interested in estimating certain parameters of a defined population or testing hypotheses on them. A clear understanding of the objectives of the study will influence the parameter(s) of importance. Consider the situation of environmental pollution in a certain city or region. Suppose that a measure of environmental pollution is the level of carbon monoxide in the air. A parameter could be the mean level of carbon monoxide per specified unit volume. This entry describes the types and categories of parameters; discusses the use of parameters in probability distributions, estimation, and hypothesis testing; and suggests a focus for future research.

## Types of Parameters

In general, parameters are either discrete or continuous. For the example considered previously (mean level of carbon monoxide per unit volume), the parameter is continuous. It can take on an infinite number of values on a continuum scale. On the other hand, consider a binomial distribution, where each person is asked whether he or she prefers a given product or not. In this situation, there are two parameters. The first represents the number of persons selected ($n$), and the second is the proportion of people who prefer the product ($p$). The parameter $n$ is discrete, whereas the parameter $p$ is continuous.

Different categories of parameters may define different features associated with the population. Consider, for example, the location parameter—three common ones are mean, median, and mode. The population mean, $\mu$, is a measure of location or central tendency of the elements in the population. It is one of the most widely used parameters of interest in research designs. Typically, when an experiment is conducted based on a design, one is interested in determining if a treatment has any effect on the population mean response. The median could be of interest in applications of nonparametric tests, where the assumption of normality of the response variable may not hold. The mode is used very infrequently in research design as a parameter.

Another category of parameters is that associated with variability in a population. Examples are the range, standard deviation, variance, and interquartile range. Of these, the standard deviation and variance are widely used. The population standard deviation is usually denoted by $\sigma$ and the population variance by $\sigma^2$. In a majority of research design problems, even if the population dispersion parameter, $\sigma^2$, is not directly of interest, it does show up indirectly in assumptions pertaining to statistical hypothesis tests on the population means. A common assumption in these tests is that of homogeneity of population variances, which assumes that the variances of the population distributions associated with each treatment are equal.

## Probability Distribution Parameters

Parameters that completely specify the probability distribution of a certain random variable are known as probability distribution parameters. The most commonly used one in practice relates to the normal probability distribution, whose density function is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right]. \qquad (1)$$

The two parameters in Equation 1 are the population mean, $\mu$, and the population standard deviation, $\sigma$. Knowledge of these two parameters gives us complete information about the nature of the distribution. Hence, it is of interest to estimate these parameters or test hypotheses on them.

## Estimation and Hypothesis Testing on Parameters

In general, point estimation and interval estimation techniques exist. In point estimation, a single number or value is used as an estimate, while in interval estimation a range or interval is specified, such that the probability of the interval containing the desired parameter, is of a desired probability level. Concepts of interval estimation are based on the sampling distribution of the associated estimator.

A common estimator of the parameter population mean, $\mu$, is the sample mean, $\overline{X}$. Through use of the Central Limit Theorem, the sampling distribution of the estimator may be established. Similarly, the sample variance, $s^2$, is a common estimator of the population variance, $\sigma^2$, where $s^2$ is given by

$$s^2 = \sum \frac{(X_i - \overline{X})^2}{(n-1)}, \tag{2}$$

where $n$ represents the number of observations in the sample, $X_i$ represents the $i$th observation, and $\overline{X}$ represents the sample mean. Note that $\overline{X}$ is obtained from

$$\overline{X} = \frac{(\sum X_i)}{n}. \tag{3}$$

There are two desirable properties of estimators. The first refers to that of unbiasedness, whereby the expected value of the estimator equals the true value of the parameter. The second desirable property is that of minimum variance, which assures us that the estimator does not fluctuate much from sample to sample.

In research designs, a simple hypothesis or a joint hypothesis could be of interest. A simple hypothesis is used when we want to test whether the effect of a treatment achieves a certain mean response. For example, does a certain drug change the mean cholesterol ($\mu$) from a certain level ($\mu_o$)? The null and alternative hypothesis could be stated as

$$H_0 : \mu = \mu_o \tag{4}$$

$$H_a : \mu \neq \mu_o.$$

Alternatively, if several drugs are being tested on their effect on mean cholesterol levels, it might be of interest to determine if at least one drug has a significantly different effect from the others. The following hypotheses could be stated:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_p \tag{5}$$

$H_a$ : At least one $\mu_i$ is different from the others.

Here, $p$ represents the number of drugs used in the design.

## Future Research

Ever since the discovery of the Gaussian (or normal) distribution to describe certain phenomena that exist in nature, parameters have a special place in research designs. Parameters not only describe the properties and characteristics associated with the particular distribution, they assist in making predictions. However, in this process of making predictions, through experimentation, parameter values are usually unknown. Future work should focus on effective and efficient estimation methods.

*Amitava Mitra*

***See also*** Central Tendency, Measures of; Estimation; Hypothesis; Population; Variability, Measure of

### Further Readings

Cox, D. R. (2006). *Principles of statistical inference*. New York: Cambridge University Press.

Keppel, G., & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

Mendenhall, W., Reinmuth, J. E., & Beaver, R. (1993). *Statistics for management and economics* (7th ed.). Belmont, CA: Duxbury Press.

Mitra, A. (2008). *Fundamentals of quality control and improvement* (3rd ed.). Hoboken, NJ: Wiley.
Winer, B. J., Brown, D. B., & Michels, K. M. (1991). *Statistical principles in experimental design* (3rd ed.). New York: McGraw-Hill.

# Parametric Statistics

Parametric statistics are the most common type of inferential statistics. Inferential statistics are calculated with the purpose of generalizing the findings of a sample to the population it represents, and they can be classified as either parametric or nonparametric. Parametric tests make assumptions about the parameters of a population, whereas nonparametric tests do not include such assumptions or include fewer. For instance, parametric tests assume that the sample has been randomly selected from the population it represents and that the distribution of data in the population has a known underlying distribution. The most common distribution assumption is that the distribution is normal. Other distributions include the binomial distribution (logistic regression) and the Poisson distribution (Poisson regression). Additionally, parametric statistics require that the data are measured using an interval or ratio scale, whereas nonparametric statistics use data that are measured with a nominal or ordinal scale. Some frequently used parametric tests that include the assumption of a normal distribution include the following: Student's $t$ test, analysis of variance, Pearson's $r$ correlation, and linear regression. Selection of the correct statistical test is important because if the wrong test is selected, the researcher increases the chances of coming to the wrong conclusion.

This entry first reviews inferential statistics methodologies. Next, it discusses the underlying assumptions for parametric statistics and parametric methods, and corrections for assumption violations. Finally, this entry discusses the advantages of parametric statistics.

## Inferential Statistics

There are two types of methodologies used in inferential statistics: hypothesis testing and estimation of population parameters. Each of these methodologies includes parametric and nonparametric tests.

Hypothesis testing uses sample data to test a prediction about a population or the relationship between two or more populations. The predictions are stated as two statistical hypotheses. The null hypothesis, $H_0$, states that there is no effect or no difference. The alternative hypothesis, $H_a$, indicates the presence of an effect or difference. This is usually the hypothesis that the researcher expects to be supported. Within hypothesis testing, it is possible for a researcher to commit two types of errors, Type I and Type II. A Type I error occurs when a true null hypothesis is rejected; that is, the conclusion is made that the alternative hypothesis is true when it is not. The likelihood of committing such an error is specified by the alpha level. For example, the likelihood of committing a Type I error if $\alpha = .05$ is 5%. A Type II error occurs when a false null hypothesis is not rejected; that is, the conclusion is made that an alternative hypothesis is not true when it is. The likelihood of committing such an error is specified by beta, $\beta$. This is related to the power of a statistical test. Power refers to the probability that a null hypothesis will be rejected when it is false—in other words, the probability of finding a statistically signi-ficant result. Power depends on the significance level ($\alpha$), the sample size, and the population effect size.

Estimation of population parameters includes point estimation and interval estimation. Point estimation involves estimating the parameter value from the computed sample statistics. It is the "best guess" for an unknown population parameter. Statistics refer to characteristics of an observed sample and are measured using measures of central tendency (i.e., mean, median, mode) and measures of variability (i.e., variance and standard deviation). Parameters refer to characteristics of a population. Interval estimation, the more commonly used, involves using sample data to compute a range of values that includes the unknown population parameter. The confidence interval is the most common form of interval estimation.

An additional measure that is often estimated is the effect size, or magnitude of treatment effect. This value estimates the strength of an apparent relationship by estimating the proportion of variance on the dependent variable that can be

attributed to the variance of the independent variable, rather than assigning a significance value.

## Assumptions

The term *parametric* was first introduced by Jacob Wolfowitz in 1942 as the opposite of *nonparametric*. He described parametric statistics as those that assume variables have a known functional form and nonparametric as those where the functional form is unknown. Parametric statistics allow one to use sample statistics to make inferences about a population parameter or parameters. Nonparametric tests, on the other hand, make fewer assumptions about the underlying distribution and therefore do not allow one to make inferences about the population. Assumptions may be thought of as "rules" or "guiding principles" for a particular statistic. For example, in analysis of variance (ANOVA), three assumptions must be met for this parametric test to be used. The first assumption is that the data follow a normal distribution. The second assumption is that the variances between groups are equal. This assumption is referred to as the homogeneity of variance assumption. The final assumption of ANOVA assumes that errors associated with any pair of observations are independent (either within the same group or between groups). This assumption is known as the independence of errors assumption, and the violation of this assumption is typically related to design. It is important to recognize and adhere to the assumptions of the test. Violations of assumptions of parametric tests decrease the reliability of the test statistic. However, many parametric tests are considered robust to certain violations—that is, the Type I error rate is not affected despite departures from assumptions. For example, ANOVA has been shown to be somewhat robust to departures from normality but is less robust to violations of homogeneity of variance.

### *Distribution*

As previously mentioned, the most common assumption of parametric tests is that the population from which the sample is taken has an underlying normal distribution. Knowing the underlying distribution of the variable allows for predictions to be made. That is, with repeated sampling of equal sizes of a particular population, the estimate will be approximately the same. The shape of the normal distribution is such that the closer a score is to the mean, the more frequently it occurs.

### *Sample*

Generally, it is not possible to collect data on an entire population; therefore, sample data are used in order to make inferences to the population. It is important to have a representative sample of a population in order to make generalizations to that population. Ways to ensure a representative sample include random sampling and sampling as large of a sample as is feasible. Having a random sample is an assumption of inferential statistics as it decreases the odds of having a biased sample. Additionally, the larger the sample size, the more likely it is that the shape of the sampling distribution will approach a normal distribution. This principle is known as the central limit theorem.

### *Measurement Issues*

Another assumption of parametric tests is based on the level of measurement. The level of measurement is a representation of how much information the data are measuring. Parametric tests require that the data include more information and thus require interval- or ratio-scaled data. An interval scale is characterized as having an equal distance between measurements throughout the scale that correspond to equal differences in the quality being measured that can be compared in a meaningful way. Interval scales do not have a true zero. A ratio scale has the same equal and meaningful distances between measurements as the interval scale and does have a true zero. Data that follow a nominal or ordinal scale contain less information and are typically analyzed using nonparametric tests. Some controversy exists in the categorization of measurement levels and their application to the decision of whether it is appropriate to use parametric statistics, particularly in determining whether a particular test is on an interval scale (parametric statistics may be used) or an ordinal scale (nonparametric is preferred).

## Parametric Methods

### *Single Sample*

Parametric tests that are concerned with testing a hypothesis about a population mean include the single-sample $z$ test and single-sample $t$ test. Parametric tests for population parameters other than the mean include single-sample chi-square test for population variance, single-sample tests for evaluating population skewness and kurtosis, and the D'Agostino-Pearson test of normality.

### *Independent Samples*

For hypothesis testing about the difference between two independent population means, the $t$ test for independent samples may be used. For multiple groups, one could use single-factor between-subjects analysis of variance (ANOVA), single-factor between-subjects analysis of covariance (ANCOVA), or multivariate analysis of variance (MANOVA).

### *Dependent Samples*

For hypothesis testing about the mean differences between two variables measured in the same sample, the $t$ test for dependent samples or the $z$ test for two dependent samples may be used. Additionally, if more than two variables are measured in the same sample, a within-subjects ANOVA, ANCOVA, or MANOVA may be used.

### *Relationships Between Variables*

To express a bivariate relationship, a Pearson product-moment correlation coefficient, Pearson's $r$, may be calculated. Multivariate extensions include the multivariate correlation coefficient, partial correlation coefficient, and semipartial correlation coefficient.

### *Effect Size Estimates*

Parametric effect size estimators are associated with the particular statistical test being utilized and include the Pearson's $r$ correlation, Cohen's $d$ index, Cohen's $f$ index, Hedges' $g$, Glass's $\Delta$, and omega squared.

## Corrections for Assumption Violations

As mentioned, the majority of parametric statistical tests include the assumption of normality. It is important to investigate one's sample data to ensure that this assumption has been met, and if it has not, there are possible ways to correct the violation. Significance testing for skewness and kurtosis can be used to test for normality. Non-normally distributed data can occur for multiple reasons. One reason may be due to data that were entered incorrectly in the data analysis program or the inclusion of incomplete data in the analyses. This problem can be remedied by reviewing the data in the data analysis program and correcting the incorrect value or by adjusting for incomplete data in the analyses.

### *Handling Outliers*

Another possible reason for non-normally distributed data is the presence of outliers. Outliers are extreme values on a variable (univariate outlier) or an extreme combination of values on two or more variables (multivariate outlier). Outliers are problematic as they influence results more than other points, and they tend to be sample specific, thus hurting the generalizability of results across samples and therefore the ability to make inferences to the population. A univariate outlier can be identified by investigating a frequency distribution table and identifying values that meet a particular criterion, such as mean $\pm 3$ standard deviations or median $\pm 2$ interquartile ranges. A multivariate outlier may be identified by examining scatterplots or using Mahalanobis distance. There are two options for dealing with outliers. If it can be determined that the outlier is not from the intended population, it may be deleted. If it cannot be determined that the outlier is not from the intended population, the outlier may be changed so as to have less impact on the results, such as bringing the outlier to the upper or lower criterion used for identification.

### *Data Transformation*

Another possible solution for dealing with data that violate an assumption of a parametric test is to transform the data so that they fit the assumption more satisfactorily. Data transformations put

the data on another scale, which can help with outliers and deviations from normality, linearity, and homoscedasticity. Particular transformations can be more beneficial depending on the types of skew. For example, when the data are positively skewed, a square root transformation can often produce a more normally distributed pattern of data or even equate group variances. For a severe positive skew, logarithm transformation may be useful; for negative skew, reflecting the variable and then using a square root transformation, and for severe negative skew, reflecting the variable and then using a logarithm transformation are useful. After the transformation, it is important to recheck the assumptions. Although data transformation can help, it may also make data interpretation difficult or unclear.

### *Robust Statistics*

Robust statistics may also be of use when one needs to compensate for violated assumptions. For example, Welch's *t* test is used when the homogeneity of variance assumption is violated and provides a corrected critical value. The term *robust* means that even when an assumption is violated, the Type I error rate will remain within an acceptable range (typically .05). However, when choosing a robust statistic, one should keep in mind that many are considered robust only when a single assumption has been violated, as in the case of Welch's *t* test.

Finally, increasing the sample size by collecting additional data may also remedy any violations to key assumptions.

## Advantages

Over the years, there has been considerable debate over whether to use parametric or nonparametric tests. In general, most researchers agree that when the data are interval or ratio and one has no reason to believe that assumptions have been violated, parametric tests should be employed. However, if one or more assumptions have been violated, nonparametric methods should be used. It is preferable to use nonparametric tests in these cases, as violations of assumptions of parametric tests decrease the reliability of the test statistic. Additionally, if the sample size is sufficiently large, it is appropriate to use parametric statistics as the larger the sample size, the more normal the sampling distribution. However, some researchers believe that nonparametric tests should be the first choice, as it is often difficult to show that assumptions have not been violated, particularly assumptions of normality.

One advantage of parametric statistics is that they allow one to make generalizations from a sample to a population; this cannot necessarily be said about nonparametric statistics. Another advantage of parametric tests is that they do not require interval- or ratio-scaled data to be transformed into rank data. When interval- or ratio-scaled data are transformed into rank-scaled data, it results in a loss of information about the sample. In practice, many researchers treat variables as continuous, and thus use parametric statistical methods, when the underlying trait is thought to be continuous even though the measured scale is actually ordinal. This is particularly true when the number of categories is large, approximately seven or more, and the data meet the other assumptions of the analysis. A third advantage of using parametric tests is that these tests are able to answer more focused research questions. For example, the nonparametric Whitney-Mann-Wilcoxon procedure tests whether distributions are different in some way but does not specify how they differ, whereas the parametric equivalent *t* test does. Additionally, because nonparametric tests do not include assumptions about the data distribution, they have less information to determine significance. Thus, they are subject to more Type II error. Therefore, if a parametric test is appropriate, it gives one a better chance of finding a significant result.

One practical advantage of parametric tests is the accessibility of statistical software programs for conducting parametric tests. There are numerous software programs readily available for conducting parametric tests, whereas there are very few that produce confidence intervals for nonparametric tests. In addition, the computational fomulas for conducting parametric tests are much simpler than those for nonparametric tests.

When choosing whether to use parametric or nonparametric tests, many researchers consider parametric tests to be robust and do not

recommend replacing them unless the violations of normality and homogeneity of variance are severe.

There are many advantages to using parametric statistics as opposed to nonparametric statistics. However, deciding which type of inferential statis-tic (parametric vs. nonparametric) is most appropriate given the particular situation is a decision that must be considered and justified by the researcher.

*Patricia Thatcher Kantor and Sarah Kershaw*

***See also*** Analysis of Variance (ANOVA); Central Limit Theorem; Correlation; Effect Size, Measures of; Levels of Measurement; Nonparametric Statistics; Normality Assumption; Normalizing Data; Outlier; Regression to the Mean; Robust; Student's *t* Test

**Further Readings**


Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Feinstein, C. H., & Thomas, M. (2002). *Making history count: A primer in quantitative methods for historians*. Cambridge, UK: Cambridge University Press.

Keren, G., Lewis, C. (1993). *A handbook for data analysis in the behavioral sciences: Methodological issues*. Hillsdale, NJ: Lawrence Erlbaum.

Maxwell, S. E., Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

McNabb, D. E. (2004). *Research methods for political science: Quantitative and qualitative methods*. New York: M. E. Sharpe.

Osborne, J. W. (2002). Notes on the use of data transformations. *Practical Assessment, Research Evaluation, 8*(6). Retrieved from http://pareonline.net/getvn.asp?v = 8&n = 6

Osborne, J. W., Overbay, A. (2004). The power of outliers (and why researchers should always check for them). *Practical Assessment, Research Evaluation, 9*(6). Retrieved from http://pareonline.net/getvn.asp?v = 9&n = 6

Sheskin, D. J. (2007). *Handbook of parametric and nonparametric statistical procedures*. Boca Raton, FL: Chapman Hall/CRC.

Tabachnick, B. G., Fidell, L. S. (1996). *Using multivariate statistics*. New York: HarperCollins College.


# Partial Correlation

A partial correlation is a measure of the relationship that exists between two variables after the variability in each that is predictable on the basis of a third variable has been removed. A partial correlation, like a conventional Pearson product-moment correlation, can range from −1 to +1, but it can be larger or smaller than the regular correlation between the two variables. In fact, a partial correlation is simply a conventional correlation between two sets of scores. However, the scores involved are not the scores on the original variables but instead are residuals—that is, scores reflecting the portion of the variability in the original variables that could *not* be predicted by a third variable.

As a concrete example, a researcher might be interested in whether depression predisposes individuals to interpret neutral faces as manifesting a negative emotion such as anger or sadness, but may want to show that the relationship is not due to the effects of anxiety. That is, anxiety might be confounded with depression and be a plausible rival explanatory variable, in that more depressed individuals might tend to have higher levels of anxiety than less depressed individuals, and it may be known that more anxious individuals tend to interpret emotions in neutral faces negatively. A partial correlation would be directly relevant to the researcher's question and could be used to measure the relationship between depression and a variable assessing the tendency to interpret neutral faces negatively, after one has removed the variability from each of these variables that is predictable on the basis of an individual's level of anxiety.

Partial correlations are convenient and have wide applicability, but they are often misunderstood. There are conceptual and statistical issues that must be kept in mind when working with partial correlations. The fomula used in the computation of a partial correlation is first introduced, and then several issues bearing on the interpretation of partial correlations are briefly considered.

## Fomula for Partial Correlation

The partial correlation, denoted $r_{12.3}$, between a pair of variables (Variables 1 and 2) controlling for a "third" variable (Variable 3) may be readily computed from the pairwise or bivariate correlations among the three variables by using the fomula

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}.$$

Understanding partial correlations is greatly facilitated by realizing how this fomula flows from the definition of a correlation. A correlation may be defined either as the ratio of the covariance between two variables to the product of their standard deviations, or as the average product of the individuals' $z$ scores or standard scores on the two variables:

$$r_{XY} = \frac{s_{XY}}{s_X s_Y} = \frac{\sum z_{X_i} z_{Y_i}}{N}.$$

(The final value of $r$ is unaffected by whether $N$ or $N - 1$ is used in the denominators of fomulas for the covariance and variances because the sample size terms in the numerator and denominator cancel out. This may be less obvious but is still true in computing $r$ as the average product of $z$ scores, given the standard deviations used to compute the $z$ scores could have been calculated using either $N$ or $N - 1$. $N$ is used on the right in the equation to conform to the typical connotation of computing the "average" product of $z$ scores as implying a division by $N$, although it should be noted that this implies that the "average" or "standard" deviations also are computed using $N$ rather than $N - 1$.)

As noted, a partial correlation is a correlation between residuals, or, as they are sometimes termed, partialled scores. These can be expressed most succinctly when the original variables are put in standard score or $z$ score form. In the example cited, if depression is denoted as Variable 1, negativity in reaction to faces as Variable 2, and anxiety as Variable 3, one might denote the partialled depression score as $z_{1-3}$, which is the residual score computed as the difference between an individual's original depression score and the level of depression predicted on the basis of the individual's anxiety. The prediction is done by linear regression; with variables in standard score form, the predicted score, denoted $\hat{z}$, for a variable is simply the $z$ score on the other variable times the correlation between the two variables. That is,

$$\begin{aligned}\text{depression partialling anxiety} = z_{1-3} &= z_1 - \hat{z}_1 \\ &= z_1 - r_{13}z_3,\end{aligned}$$

where $z_1$ = standard score on depression; $z_3$ = standard score on anxiety; $r_{13}$ = correlation between depression and anxiety; $z_{1-3}$ = residual score on depression after variability associated with anxiety is removed.

Similarly, the residual score on negativity in reaction to faces would be

$$\begin{aligned}\text{negativity partialling anxiety} = z_{2-3} &= z_2 - \hat{z}_2 \\ &= z_2 - r_{23}z_3.\end{aligned}$$

These residual scores are uncorrelated with anxiety, and hence their correlation will be unaffected by any linear relationship with anxiety. We can now develop the fomula for the partial correlation by substituting these residuals into the conventional fomula for the correlation:

$$r_{12.3} = r_{(1-3)(2-3)} = \frac{s_{(1-3)(2-3)}}{s_{(1-3)}s_{(2-3)}}.$$

The denominator terms indicate the variability of the residuals. But note that because the squared correlation between two variables indicates the proportion of variability accounted for, and because the variance of the standardized variables is 1, the standard deviation of each set of residuals is simply the square root of 1 minus the squared correlation between the two variables:

$$s_{(1-3)} = \sqrt{1 - r_{13}^2} \text{ and } s_{(2-3)} = \sqrt{1 - r_{23}^2}.$$

The covariance of the residuals in the numerator is the average of the cross products of the residual scores, which can be written as

$$
\begin{aligned}
s_{(1-3)(2-3)} &= \frac{\sum z_{(1-3)} z_{(2-3)}}{N} \\
&= \frac{1}{N} \sum (z_1 - r_{13} z_3)(z_2 - r_{23} z_3) \\
&= \frac{1}{N} \sum (z_1 z_2 - r_{13} z_2 z_3 \\
&\quad - r_{23} z_1 z_3 + r_{13} r_{23} z_3^2) \\
&= r_{12} - r_{13} r_{23} - r_{23} r_{13} + r_{13} r_{23} \\
&= r_{12} - r_{13} r_{23}.
\end{aligned}
$$

Substituting these values into the numerator and denominator of the fomula for the correlation, we obtain as desired

$$
\begin{aligned}
r_{12.3} = r_{(1-3)(2-3)} &= \frac{s_{(1-3)(2-3)}}{s_{(1-3)} s_{(2-3)}} \\
&= \frac{r_{12} - r_{13} r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}.
\end{aligned}
$$

Thus, we have shown that the partial correlation is the ratio of a numerator indicating the covariance of the residual scores from which the third variable has been partialled to a denominator that is the product of the standard deviation of these residual scores.

## Issues in Interpreting Partial Correlations

There are several reasons why a partial correlation may not be meaningful or its meaning might be misunderstood. Thus, conceptual and statistical issues that complicate the interpretation of partial correlations must be considered.

First, the residual score computed may not capture the construct of interest. One may, for example, conceive of depression as sharing with anxiety certain psychological processes or symptoms such as negative affect. Removing variability from depression associated with anxiety might then remove core aspects of what constitutes depression, so that the residual scores are, in fact, considerably less valid than the original scores as a measure of depression. A related issue arises when one of the variables is a discrete variable indicating group membership such as race. One might be interested in knowing whether racial groups would differ in some outcome such as academic achievement if they did not differ in socioeconomic status (SES). Using a partial correlation to shed light on this may be helpful, but it must be kept in mind that the question posed is a counterfactual one (i.e., the facts may be that there are differences across racial groups in SES), and the answer provided implicitly assumes that the relationship between SES and achievement is a strictly causal one, which it may not be. Thus, a partial correlation that estimates the relationship between achievement and racial group membership controlling for SES may not correspond to reality, just as comparing high and low depression groups controlling for anxiety may not be representative of actual groups of depressed and nondepressed people.

Second, a heuristic principle often suggested for understanding the meaning of partial correlations may not apply in a researcher's real-world application. The heuristic often suggested to aid in understanding the meaning of a partial correlation is that it reflects the correlation expected between two variables when a third variable is held constant. The key assumption that assures this heuristic actually holds is multivariate normality. Multivariate normality implies not only that each of the three variables involved has a univariate normal distribution, but also that the joint density function of each pair of variables must be bivariate normal. If this is not the case, then the conditional correlation (arrived at by pooling the correlations between the two variables computed within each of the various levels of the third variable) may not correspond to the partial correlation. As a striking example of this, consider the relationship between statistical independence and zero correlation that obtains in the two-variable situation and how this relates to the facts of partial correlations. That is, it is well-known that a zero correlation between two variables does not imply their statistical independence because of the possibility of nonlinear relationships, but it is also well-known that if two variables are statistically independent, they must have a zero correlation. In sharp contrast, it is possible that two variables might be conditionally independent at each level of a third variable, yet their partial correlation controlling for the third variable may be nonzero. This can arise, for example, when there is a nonlinear relationship between the third variable and one of the other variables.

Thus, a third point to keep in mind in working with partial correlations is that conventional partial correlations control only for linear relationships among variables. If one suspects nonlinearity, higher-order partial correlations may be appropriate. A higher-order partial correlation is computed by statistically controlling for multiple variables simultaneously. In such a situation, you are still computing a correlation between residuals, but the residuals reflect the variability in a variable that is not predictable by any of the variables being controlled for. To extend our original example, one might compute the partial correlation between depression (Variable 1) and negativity (Variable 2), controlling for both anxiety (Variable 3) and gender (Variable 4), with the higher-order partial correlation being denoted $r_{12.34}$. Thus, one way of dealing with suspected nonlinear relationships between a third variable and one or both of the variables of most interest would be to compute a higher-order partial correlation in which one partialled out various powers of the third variable (for example, one might control for not only the anxiety score but also anxiety squared).

Finally, in computing partial correlations, one is removing the effects of a construct only to the extent that one has validly and reliably measured that construct. If one's measure of anxiety is poor, then some effects of anxiety will remain in the variables of primary interest even after the poor measure is controlled for. One common example of such fallibility is when variables are used in research that have been formed by artificially dichotomizing continuous variables. In such a case, the dichotomous variable will have considerably less information about the construct of interest than the original continuous form of the variable. Controlling for such an imperfect measure of a construct of interest can lead to spurious conclusions and result in either positively biased or negatively biased estimates of the strength of the partial correlation between the other variables.

### Conclusion

A partial correlation is a potentially very useful measure of the relationship between two variables that remains after the linear relationship of each of the variables with a third variable has been removed. However, because of the potential effects of factors such as nonlinearity, unreliability, and nonnormality, partial correlations must be interpreted with caution.

*Harold D. Delaney*

***See also*** Analysis of Covariance (ANCOVA); Confounding; Correlation; Covariate; Multivariate Normal Distribution; Regression to the Mean; Standard Deviation; *z* Score

#### Further Readings

Maxwell, S. E., & Delaney, H. D. (1993). Bivariate median splits and spurious statistical significance. *Psychological Bulletin, 113,* 181–190.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Vargha, A., Rudas, T., Delaney, H. D., & Maxwell, S. E. (1996). Dichotomization, partial correlation, and conditional independence. *Journal of Educational and Behavioral Statistics, 21,* 264–282.

Zinoverlineg, R. E., Suzuki, S., Uliaszek, A. A., & Lewis, A. R. (in press). Inflated Type I error rates in analysis of covariance (and analysis of partial variance) arising from unreliability: Alternative and remedial strategies. *Journal of Abnormal Psychology.*

## Partial Eta-Squared

Partial eta-squared is an estimate of effect size reported by SPSS, an IBM company, in conjunction with analysis of variance (ANOVA) and generalized linear model (GLM) analyses. Although there is general consensus about the desirability of reporting estimates of effect size in research reports, there is debate about the relative utility of various options. Relative to other options such as eta-squared or omega squared, partial eta-squared is less well understood and less useful. The recent popularity of partial eta-squared appears to rest on its inclusion in the output of popular statistical software rather than its merits as an informative estimate of effect size.

Many social scientific journals require reporting estimates of effect size to supplement tests of statistical significance, and such reporting is considered good practice by most commentators. An effect size is a standardized estimate of the

magnitude of a difference or the degree of association that is relatively independent of sample size. Simply put, estimates of effect size are designed to reflect how strongly two or more variables are related, or how large the difference is between groups. Cohen's *d*, the correlation coefficient *r*, and the squared multiple correlation $R^2$ are examples of estimates of effect size. For ANOVA, omega squared, epsilon squared, and eta-squared are common estimates of effect size. Partial eta-squared is an estimate of effect size. This entry describes the calculation of partial eta-squared and discusses several issues associated with reporting partial estimates of effect size.

## Calculation

Partial eta-squared is calculated by dividing the sum of squares for a particular factor, which could be either a main effect or an interaction, by the sum of squares for the factor plus the sums of squares error for that factor as shown in Equation 1:

$$\text{partial } \eta^2 = \frac{SS_{\text{between}}}{SS_{\text{between}} + SS_{\text{error}}} \quad (1)$$

Partial eta-squared can also be calculated from an *F* test, as shown in Equation 2:

$$\text{partial } \eta^2 = \frac{n_1 F}{n_1 F + n_2} \quad (2)$$

Equations 1 and 2 are equivalent, and both equal the squared partial correlation when there are two levels of the independent variable.

It is important to distinguish partial eta-squared from eta-squared. Eta-squared is the sum of squares for a particular factor divided by the total sum of squares and reflects the proportion of variance explained. Eta-squared is shown in Equation 3:

$$\eta^2 = \frac{SS_{\text{between}}}{SS_{\text{total}}} \quad (3)$$

In one-way ANOVA, partial eta-squared and eta-squared will always be exactly equal. This is because there are no other factors to partial. In multi-way ANOVA, however, the two values depart, sometimes radically. Partial eta-squared is always greater than or equal to eta-squared, and the discrepancy is a function of the number of factors in the design; the effects of the other factors on the dependent variable; and if the design involves only independent groups factors, or whether some factors are repeated. Partial eta-squared will be increasingly larger than eta-squared as the number of factors in the design increases, as the effect sizes for the other factors increase, and as the proportion of factors that were repeated measures increases.

Whereas cumulated eta-squared values for a particular dependent variable can never sum to more than 1.00, partial eta-squared has no such limitation. That is because eta-squared consistently uses the total sum of squares for the dependent variable in the numerator. Every explanatory factor can account only for unique variance partitioned from the total variance in the dependent variable. This makes sense in experimental designs where components of variation are orthogonal. Partial eta-squared has only the sum of squares for the effect and the error sum of squares in the denominator. Thus, each factor has a unique denominator for the effect size calculation. Each calculation of partial eta-squared has its own upper limit that approaches 1.00 (when there is essentially no variance left as unexplained or error). Thus, the theoretical upper limit for cumulated partial eta-squared values is the number of factors, *k*. That is, if there are four factors in the design, their eta-squared values can sum to 1.00, whereas their partial eta-squared values can sum to 4.00.

Partial eta-squared is sometimes misreported as eta-squared. Examples of misreporting are obvious when the effect sizes labeled as eta-squared account for a percentage of variance in a dependent variable exceeding 100%. This is not possible when calculating eta-squared, but can occur when using partial eta-squared. This reporting error can result from a problem with some versions of SPSS incorrectly labeling partial eta-squared as eta-squared. The problem, however, can also occur with other statistical software. Thus, the problem does not appear to be specific to a particular field of research nor a particular software package.

## Issues Concerning Reporting Partial Estimates of Effect Size

One key to understanding the comparative desirability of reporting partial versus nonpartialled

estimates of effect size lies with the difference between conceptualizing design factors as inducing variance versus explaining variance in the dependent variable. If research designs differ by adding additional experimental manipulations, a case can be made that the additional factors caused increased variance in the dependent variable between those two studies. Thus, one would not expect a factor in a two-way ANOVA to account for as much variance in the dependent variable as the same factor in a one-way ANOVA because of the increased sum of squares value in the former case. Under the condition that one was convinced that the effects of additional factors caused additional dependent variable variance, it could be reasonable to use partial eta-squared as a comparable index across studies with different numbers of factors. That is, partial eta-squared is useful for comparing effects for the same factor tested across different designs when the designs differ by additional experimental inductions that cause increased total variance. However, if an attribute variable such as sex is used as a factor, it cannot be claimed to have caused differential dependent variable variance across designs. Thus, it would not be appropriate to partial that component of variance out of the effect size calculation across designs. A nonpartialled eta-squared would be advised. Overall, then, only additional induced sources of variance should be removed from the denominator when making comparisons across studies. To that end, generalized eta-squared has been proposed, which partials out only sources of induced variance. Although partialling effects for the purpose of direct, across-design comparisons is conceptually sound, in practice, it is difficult to know for sure if additional factors instill variance, explain existing variance, or both. For this reason, when partial eta-squared or generalized eta-squared is reported, it is advisable to report eta-squared as well.

A second issue relates to the conceptual meaning of partialling out orthogonal factors. A common use of a partial correlation is eliminating spurious effects attributable to other independent variables. In survey research, for example, the different predictor variables are typically correlated, and it makes sense to statistically control some independent variables. In factorial experimental designs,however, where partial eta-squared is most often reported, the other independent variables are nearly orthogonal. Partialling the effects of an orthogonal, non-control variable has a different substantive interpretation.

A third issue related to which estimate of effect size to report relates to the usefulness of each for later meta-analyses. Meta-analysis typically requires zero-order indexes of effects such as *r* or *d*. Thus, indexes that control for (i.e., partial) other factors such as partial regression coefficients and partial eta-squared are of less use in these cumulative studies. Given the tendency of partial eta-squared to substantially overestimate an effect as design complexity increases, including these effect size statistics in a meta-analysis could lead to erroneous determinations of heterogeneous effects or identification of moderators that do not exist. These problems are especially relevant when attempting to convert reported *F* values to *r*. The typical fomulas used for this conversion lead to partial correlations rather than zero-order correlations when there was more than one ANOVA factor. Thus, the effect sizes derived from *F* are of little use in meta-analysis. Craig R. Hullett and Timothy R. Levine provided a method for accurately estimating zero-order effects from data typically presented in research reports as a means for overcoming this limitation. Of course, it would still be reasonable to include partialled effect sizes if they met the previously stated condition of controlling only for additional induced variation in the dependent variable. Otherwise, partialled effect sizes have less value for current methods of meta-analysis.

A final issue relates to confidence intervals. Confidence intervals are a function of the standard error of whatever statistic is being reported. They require knowledge of the sampling distribution for the statistic. The sampling distributions of eta-squared and partial eta-squared are not equivalent. Although they do not go into detail about the differences between the distributions, one can logically infer from parallel statistics the elements that will affect the difference in standard errors. For example, a regression coefficient from a design in which there are multiple independent variables has a standard error that increases as the number of additional variables increases, as the correlation among the predictors increases, and as the total explained variance by all included variables decreases. One

would expect similar consideration when determining the standard error and, thus, the confidence intervals around partial eta-squared. To date, though, no published information appears to exist detailing how such elements should be considered in calculating confidence intervals around partial eta-squared. Without such knowledge, one cannot come to a clear conclusion about whether a partial eta-squared should be considered a good or poor estimate of effect in the population.

### Final Thoughts

Partial eta-squared is an increasingly popular estimate of effect size reported by popular statistical software programs in conjunction with ANOVA and GLM analyses. The purpose of partial eta-squared is to provide an estimate of effect size for comparing the effects of a factor across studies with different research designs. Its usefulness, however, is limited by several considerations. Therefore, it should be reported as a supplement to, but not a substitute for, eta-squared or omega squared.

*Timothy R. Levine and Craig R. Hullett*

***See also*** Analysis of Variance (ANOVA); Effect Size, Meaures of; Eta-Squared; Meta-Analysis; Omega Squared; Partial Correlation

### Further Readings

Hullett, C. R., & Levine, T. R. (2003). The overestimation of effect sizes from *F* values in meta-analysis: The cause and a solution. *Communication Monographs, 70,* 52–67.

Levine, T. R., & Hullett, C. R. (2002). Eta-squared, partial eta-squared, and misreporting of effect size in communication research. *Human Communication Research, 28,* 612–625.

Murray, L. W., & Dosser, D. A. (1987). How significant is a significant difference? Problems with the measurement of magnitude of effect. *Journal of Counseling Psychology, 34,* 68–72.

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods, 8,* 434–447.

Pierce, C. A., Block, R. A., & Aguinis, H. (2004). Cautionary note on reporting eta-squared values from multifactor ANOVA designs. *Educational and Psychological Measurement, 64,* 916–924.

## Partially Randomized Preference Trial Design

Partially randomized preference trials (PRPTs), using Brewin and Bradley's design, are a product of combining the best elements of randomized controlled trials (RCTs), which involve random allocation of different treatments to willing patients, and feasibility studies, in which patients choose their preferred treatment. PRPTs give patients, who are recruited into a clinical trial, the option to choose their preferred method of treatment, and if the patients have no strong motivation toward a specific treatment, they are asked if they will agree to random allocation to one or another treatment method. All patients recruited into the PRPT need to be given clear, accurate, and detailed information about what the treatments to be offered in the trial involve. They can then make an informed decision when given the opportunity in the PRPT to choose a preferred method of treatment. PRPTs can be used to evaluate many different types of treatment, including medical treatment, psychological treatment, and dental treatment, or a combination of treatment types, for example, drug versus psychological treatment for depression.

This entry first details the structure of a PRPT and variations of the structure. Next, this entry discusses validity (external and internal), the acceptability of PRPT to patients, and PRPTs' advantages and limitations. Last, this entry describes the appropriate implementation of PRPTs.

### Structure

In a PRPT comparing two treatments, there are potentially four separate groups of patients, each receiving one of the two treatments, usually an established treatment and a new treatment. Patients are informed about the two treatments being compared and asked if they have a strong preference. Patients who have strong preferences for one treatment over another are allocated to a group in which they can have their preferred

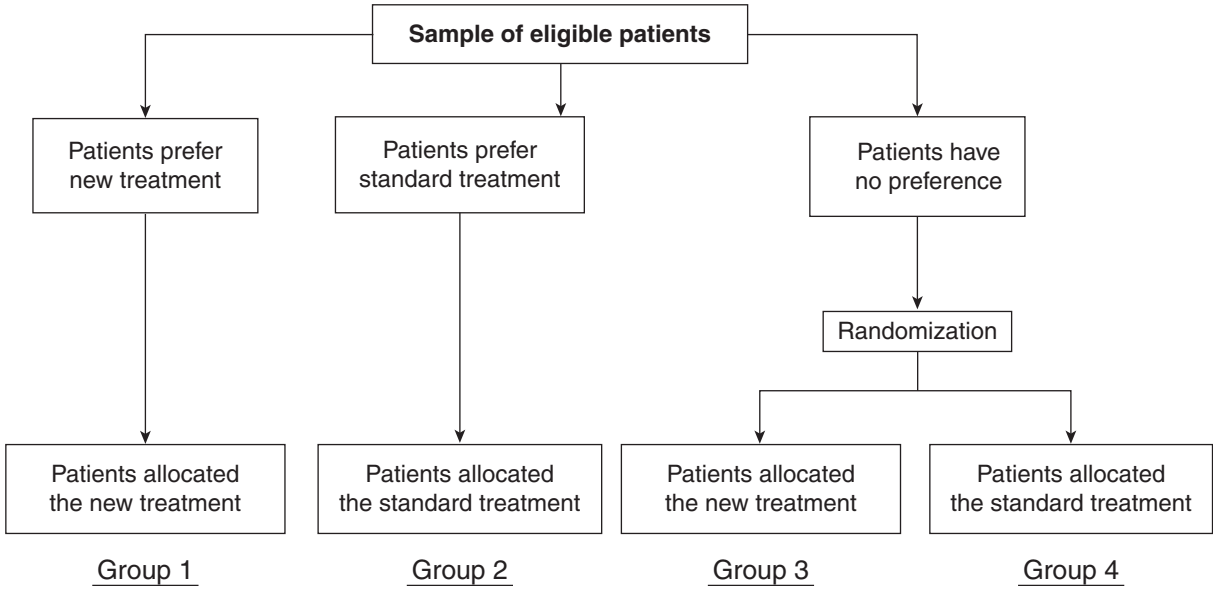



**Figure 1** The Structure of a Partially Randomized Preference Trial (PRPT)

method of treatment. In Figure 1, patients who are particularly motivated toward the new treatment are allocated to Group 1, and patients who would prefer to use an established method of treatment are allocated to Group 2. Patients who have no strong preference and are equally prepared to use either treatment are said to be in equipoise and, with their consent, have one or the other treatment type assigned randomly and are thereby allocated to either Group 3 or Group 4.

## Variations on the Structure

Three-group trials may result if no one has a strong preference for one of the treatments. A very small number of patients in one of the preference groups (Groups 1 or 2 in Figure 1) may not be analyzable statistically but will still serve an important purpose in removing those with preferences from Groups 3 and 4, where randomization to a nonpreferred treatment would lead to disappointment.

If all the patients who are recruited into the PRPT have a strong preference, a feasibility study will result where all patients chose their treatment, and if no recruit has a strong preference, an RCT will result. However, if an RCT results from patients not having a preference for a particular treatment offered, it will differ from that of many conventional RCTs as motivational factors will not distort outcomes. In a conventional RCT, some patients may agree to randomization with the hope of obtaining a new and/or otherwise inaccessible therapy. Patients who think they would prefer a new treatment (those who are allocated to Group 1 in a PRPT) may have been included in an RCT if the new treatment was unavailable outside the trial and participation in the trial was their only way of obtaining their preferred method of treatment. However, RCT participants are asked to accept any of the health care options being compared. Inclusion of patients preferring the new treatment over a standard treatment will bias the RCT sample in favor of the new treatment; those with preferences for the standard treatment are more likely to decline to participate in an RCT as they can usually obtain their preferred treatment outside the trial. The RCT sample recruited will be randomized to two groups. When preferences for the new treatment are marked, one group will contain participants who are pleased to receive the new treatment whereas the other group will contain individuals who are disappointed that they have not been allocated to the

new treatment (as demonstrated empirically by Feine and colleagues in 1998 and discussed by Bradley in a commentary the following year). When the majority of recruits have a preference for the new treatment, an RCT creates groups that have been allocated at random but differ in respect to motivation to use the treatment assigned. The control group of participants who have been randomized to the standard treatment will contain patients who are disappointed with the treatment allocation and therefore will be more likely to drop out of the trial. They may also be less likely to follow the treatment recommendations and do less well with the treatment than they would have done if motivated to use that treatment. If disappointed patients drop out of the control group, outcomes of the control treatment will be artificially improved, minimizing any advantages of the new treatment. However, if such poorly motivated patients remain in the control group, outcomes will be worsened, thereby exaggerating the advantages of the new treatment.

## External Validity

People who refuse to take part in an RCT are more likely to have strong preferences, and some participants may have strong opinions regarding the relative acceptability of the treatments being compared. People refusing to take part in an RCT because they do not want to risk being randomly allocated to a nonpreferred treatment are more likely to take part in a PRPT. Few patients invited to participate in PRPTs decline to do so (e.g., only 3 of 373 women invited declined to participate in Henshaw and colleagues' 1993 study using the PRPT design). Participants who accept an invitation for an RCT may be a minority of those invited—sometimes fewer than 20%. Within RCTs, recruitment success can vary markedly from center to center, as in the World Health Organization trial of continuous subcutaneous insulin infusion pumps versus intensified injection regimens for Type 1 diabetes in the early 1980s, where recruitment ranged from 70% in Albania to 20% in Paris. An RCT may fail to recruit patients who prefer a treatment available outside the trial, and for those who are recruited, randomization may create differences between groups; one being disappointed whereas the other is pleased with the treatment allocated at random. Patients recruited to a PRPT, on the other hand, are likely to be the vast majority of those invited, and the results obtained will therefore be more generalizable to the population of eligible patients. Patient preferences act to reduce the recruitment of patients into RCTs (as patients fear randomization will allocate a nonpreferred treatment) and thus reduce external validity of RCTs. External validity is improved by using a PRPT design, where patients know any strong preference will be met, and hence recruitment is usually close to 100%.

## Internal Validity

Not only is recruitment success greater in PRPTs compared with RCTs, but dropout rates of PRPTs are also lower in both the randomized and preference arms than are commonly found in RCTs. As more participants drop out of RCTs, the randomization process is increasingly undermined and the groups are less comparable. With PRPTs, the randomized groups have been cleared of preferences prior to randomization, and the preference groups, being self-selected anyway, are not threatened by any few dropouts that may occur.

The importance of PRPTs and patient choice of treatment are especially apparent when considering the best treatment for chronic illnesses, such as diabetes, where patient motivation is essential and factors such as the convenience and flexibility of treatment regimens can dramatically improve quality of life as seen in the DAFNE (Dose Adjustment for Normal Eating) trial. Motivation is more likely to influence outcome when there is a greater need for patient participation in their treatment, which may involve diet adjustment, self-monitoring, and self-medication regimens. When patients play such an active role in treatments and treatment impacts on lifestyle, results may be misleading if conclusions are drawn from patients who were allocated to a method of treatment they did not want (i.e., in many RCTs). Psychological and biomedical outcomes of treatment are likely to be improved when patients' preferences are matched to treatment, especially when the biomedical outcomes of the treatment are dependent upon participative interventions.

## Acceptability to Patients

The PRPT method improves treatment outcomes by increasing the chances of meeting the needs of individual patients by establishing which treatment is most suitable for a particular patient in his or her individual circumstances. Differing needs, priorities, and motivation are variables that can influence individual patient choices for medical and other treatments. There are also demographic, clinical, and biochemical variables that may influence which type of treatment is best suited to different patient subgroups.

People who design clinical trials are increasingly showing greater interest in patients' views and the impact of treatments on quality of life and other patient-reported outcomes. Such outcomes are usually facilitated by matching preferences with treatment. Although initially the PRPT was seen as a compromise to be used when many patients could not be persuaded to accept randomization to treatments about which they had strong views, now the PRPT is more often recognized as the design of choice, superior to an RCT when patients have strong preferences, and superior to an entirely unrandomized feasibility study when at least some patients can be randomized without risk of disappointment.

## Advantages of PRPTs Compared With Feasibility Studies

In feasibility studies, patients chose their own method of treatment from two or more options offered. Various physiological, psychosocial, demographic, and other characteristics of patients can affect the type of treatment that individual patients opt for and can therefore affect treatment outcomes and conclusions drawn from the trial. For example, in an obesity trial comparing a drug treatment and a diet treatment, patients who choose the drug treatment may do so because they have tried many diets and have failed to lose weight by dieting, whereas individuals who opt for the diet treatment may not have tried to diet before and may therefore anticipate more success. Such psychosocial factors are likely to influence individuals' motivation to use a particular treatment method from the outset of treatment. Individual differences in psychosocial factors may have a marked effect on the success of each treatment for an individual patient. The motivation to follow the treatment recommendations by patients who opt for the diet treatment would be greater than that of those individuals who had little faith in dieting. This motivational effect would enhance the advantages of both treatments in a self-selected sample. However, it is difficult to draw conclusions from feasibility studies about the value of treatments to patients who have no preferences. By interviewing patients and/or eliciting questionnaire data, the reasons patients have for choosing a particular treatment can be determined in both feasibility studies and PRPTs. However, the randomized subgroups in PRPTs allow for the treatments to be compared in groups where motivational factors and other baseline characteristics can be expected to be similar before and after randomization.

In feasibility trials, treatment groups may include individuals who have a strong preference for that treatment over the alternative treatment and individuals who had no preference but had to decide on a method of treatment. Patients involved in a PRPT are not pressured to choose a particular type of treatment. Only those individuals with strong preferences are allocated to the treatment group preferred; patients who do not have strong preferences are asked if they would agree to be randomized. The randomized and preference groups of the PRPT allow comparisons to be made at baseline to look for and measure differences between randomized and nonrandomized groups using each treatment in the trial with a view to understanding any between-group differences in outcome.

## Advantages of PRPTs Compared With Randomized Controlled Trials

RCTs are currently viewed by the majority of medical researchers as the gold standard for clinical trial design, and a lack of randomization is usually seen as a weakness or flaw in a study. The assumption that RCT designs are the only acceptable way to conduct clinical trials may have led to some new treatments being used without evaluation and studies to remain unpublished because randomization was deemed to be unsuitable or unethical or

was refused by patients. These studies might take place using a PRPT where ethical approval and patient acceptance could be gained from the incorporation of an element of choice of treatment type in the design.

Randomization aims to avoid the limitations of a feasibility study such as the introduction of selection bias, which can affect comparability of treatment groups and different patient characteristics affecting treatment choice. However, most proponents of RCTs have wrongly assumed that randomization always produces comparable groups and have overlooked the fact that randomization itself can create between-group differences as a result of disappointment effects. The PRPT design aims to deal with the limitations of an RCT, optimizing motivational factors in those with strong preferences and equalizing motivational factors in the randomized groups within the PRPT. The effects of motivational factors on treatment outcome can be considered by comparing those who chose a treatment with those who were content to accept randomization.

Larger numbers of participants are likely to be recruited into a PRPT over an RCT, especially in certain areas of medicine such as gynecology and obstetrics, where preferences are strong. In a clinical trial comparing treatments for menorrhagia conducted by Cooper and colleagues, the majority of women recruited (97%) agreed to participate in the PRPT design compared with 70% who agreed to take part in the conventional RCT. The 40% relative increase in participants agreeing to participate in the PRPT suggests an increase in perceived acceptability of a PRPT over an RCT where patients are likely to have strong views concerning treatment.

It is more difficult to recruit patients with a preference for a conventional treatment readily available outside the trial. Therefore, the majority of participants in an RCT will be biased in favor of the new treatment and less likely to make a nonpreferred treatment work as well as a preferred treatment. In RCTs, people are likely to be more disappointed if allocated to the conventional treatment group, and this problem is likely to arise whenever a new treatment is only available to patients within a clinical trial. Therefore, the overall sample studied in a PRPT or in a feasibility study is likely to be more representative of clinical reality and less biased by the effects of experimental manipulation. The disappointed patients in an RCT who are randomized to the nonpreferred, conventional treatment, if they remain in the trial, are more likely to stop following the standard treatment recommendations and do less well, leading to artificially advantageous conclusions favoring the new treatment. The extent of the advantage overestimation will be dependent upon the proportion of disappointed patients who remain in the trial. If these disappointed patients drop out, outcomes will be artificially improved in the randomized conventional treatment group (Group 4 in Figure 1). Outcomes of PRPTs are likely to be more successful than those found in a conventional RCT as the patients in a PRPT with strong views are generally more likely to make their treatment work. Torgerson and colleagues recommend as an alternative to PRPTs that preferences be measured in RCTs and subgroup analyses conducted to examine the impact of preferences. However, the RCT may lose patients with the strongest preferences at least for the standard treatment and maybe for both treatments, who then won't be available for such analyses, biasing the conclusions of the RCT. Thus, although Torgerson and colleagues' approach may enhance the value of RCTs, it does not provide a substitute for PRPTs.

## Limitations

The cost difference between a clinical trial that uses an RCT design and one that uses a PRPT design is sometimes cited as a reason for opting for a conventional RCT design. More resources are required by a PRPT as twice as many groups are potentially involved in the trial compared with those included in an RCT. Patients who had a preference for the control treatment are followed up in a PRPT. These patients would probably not be recruited or followed up in an RCT. Patients with a preference for the new treatment unavailable to them outside the trial would probably be recruited into an RCT, but disappointment effects can result and distort the findings of an RCT.

The cost savings of an RCT must be weighed against the loss of internal and external validity due to having a nonrepresentative sample and outcomes that are potentially distorted by disappointment

effects. Even when disappointment effects in an RCT are minimized to those of a PRPT (i.e., when both treatments being compared are available outside the trial), the PRPT still retains the advantage of recruiting a representative sample, not just those willing to be randomized, but also those with strong preferences who would decline to participate if it meant they might be randomized to a nonpreferred treatment.

## Implementation

There are several steps that need to be followed by clinical trialists to ensure a PRPT design is implemented appropriately. Patients must be given detailed and accurate information about all the treatments that are available in the trial, including the possible side effects and likely effects of treatment on lifestyle and quality of life. Only when such information is provided can patients make judgments about which treatment, if any, they prefer. Patients should then be asked if they have a strong preference for one or another treatment *before* they are asked if they are willing to have their treatment determined at random. Patients with strong preferences are then allocated to their preferred treatment group and the remaining patients are said to be in equipoise; they have been informed of the treatments but have no clear preference and can have treatment allocated at random to create two comparable groups.

If the procedure used to recruit patients in a PRPT fails to achieve equipoise in the patients who have treatment allocated at random, the randomization process cannot create two similar groups for comparison. Consider, for example, Henshaw et al.'s article that reported a PRPT design comparing medical versus surgical abortion. The trial was presented the trial to patients in a patient information sheet as if it were a conventional RCT. If patients refused to participate, then the clinician informed them they could, in fact, have their preferred treatment. However, it is likely that not all patients with strong preferences will decline to be randomized, particularly if they believe that their preferred treatment is only available within the trial. The randomized groups may therefore have contained patients with strong preferences, and if so, those groups will not have been comparable. Disappointment effects may have distorted the findings from the randomized groups compared with the results from randomized groups including only patients in equipoise. The impact of disappointment effects in this trial of different methods of pregnancy termination is likely to have been less than would be the case for participative treatments for long-term conditions where disappointment may impact on adherence and impair biomedical outcomes as well as quality of life and other patient-reported outcomes. Despite these concerns, the procedures used by Henshaw et al. have been followed by many subsequent trialists adopting the PRPT design to study more participative treatments.

A systematic review of 32 trials that took account of preferences was reported in the *Journal of the American Medical Association* by King and colleagues in 2005. Twenty-seven of these trials used PRPT designs referred to as "comprehensive cohort trials," and five were referred to as "two-stage randomized trials" involving an initial randomization to PRPT or RCT and then subsequent randomization within the RCT or for those patients in equipoise within the PRPT. The 32 trials covering a range of clinical areas (gynecology, depression, and cancer among them) included some PRPTs that were inappropriately implemented and others where implementation was insufficiently well specified to determine appropriateness. Trialists did not usually report whether or not the treatments under investigation in the trial were available outside the trial, and it was unclear whether patients believed the treatments to be available outside the trial. If patients assumed that a new treatment under investigation was only available to them within the trial, those with a strong preference for the new treatment would have been likely to agree to take part in the trial and undergo randomization (if they were not initially asked if they had a preference) in order to have a 50% chance of obtaining the new, preferred treatment. Those with a preference for the standard control treatment may assume they could have that existing treatment outside the trial and decline to participate. This is where optimally conducted PRPTs are necessary to avoid disappointment effects and treatment-related attrition that undermine RCTs. Critiques of King et al.'s review point out that the authors overlooked the critical difference between the randomized arms of an

RCT, which, when the treatments are not both available outside the trial, may be biased by disappointment effects and dropouts, and the randomized arm of a PRPT where, when properly implemented, patient equipoise is achieved and biases eliminated.

In those trials that included a description of allocation of intervention among those reviewed by King and colleagues, the majority followed Henshaw and colleagues' example of only offering participants a choice of treatment if they refused randomization. The likely result is the inclusion of patients with preferences (who did not refuse randomization) in the randomized groups and therefore randomized groups that were not in equipoise.

The randomized groups within PRPTs should not be assumed to be the same as those in a conventional RCT. Properly conducted, the randomized groups in a PRPT will be cleared of patients with preferences, leaving only those in equipoise and removing the risk of disappointment effects that are seen in RCTs when patients have preferences. Even improperly conducted PRPTs that only offer choice when patients decline randomization will have a reduced risk of disappointment effects compared with those seen in RCTs. If one wishes to examine the impact of patient preferences on the outcome of trials, one needs to compare the randomized groups of a properly conducted PRPT (where patients have no strong preferences) with randomized groups in an RCT of the same new treatment only available within the trial and that patients are keen to use. The RCT groups will include patients with preferences, and disappointment effects will be likely in the group allocated to the nonpreferred standard control treatment. Two-stage randomized trials where patients are unaware of the initial stage of randomization to a PRPT or RCT design are ideal for comparing the randomized groups of a PRPT with those of an RCT, but differences should only be expected when at least some patients have strong preferences for one of the treatments that is only available in the trial. If treatments are believed by patients to be available outside the trial, even the randomized groups of an RCT may be free of disappointment effects as those with preferences decline randomization and seek their preferred treatment outside the trial. Proponents of PRPTs consider King et al.'s conclusions that RCTs remain the gold standard and that intervention preferences appear to have limited impact on the external or internal validity of randomized trials' to be unjustified and misleading. The preferred method of determining the impact of preferences on outcomes of randomized trials is to compare the randomized groups of a well-conducted PRPT (cleaned of preferences) and the randomized groups of an RCT where preferences remain and can be disappointed at randomization rather than the method employed by King et al., which compared preference groups assigned to their preferred treatment with randomized groups who had few, if any, preferences to be thwarted, from within the same PRPT.

## Optimal Use

In a PRPT, participants are given a choice of treatment and only those without a strong preference (i.e., in equipoise) are asked if they will be randomized—all others being given their preferred treatment. This ensures that no one of those in the randomized arms is likely to experience the disappointment effects often experienced in conventional RCTs and dropouts will be kept to a minimum. The PRPT design is particularly important in the treatment of long-term conditions, participative treatments, and treatments that have very different implications for patients (e.g., surgery vs. long-term medication). Where trial treatments are readily available outside the trial and patients know that they can have a preferred treatment if they decline to participate in a trial, the results of an RCT are likely to be similar to the results obtained from the randomized subgroups of a PRPT. However, the PRPT provides the opportunity to study the outcomes in patients who have strong preferences and are given their preferred treatment as well as studying patients in equipoise. Where a new treatment is only available within a trial and patients participate in the hope of having the new treatment, the results of an RCT are likely to be distorted by disappointment effects that can be avoided by using a properly implemented PRPT design.

*Joanna Bradley-Gilbride and Clare Bradley*

*See also* Ethics in the Research Process; Randomized Block Design

### Further Readings

Bradley, C. (1988). Clinical trials—Time for a paradigm shift? [Editorial]. *Diabetic Medicine, 5,* 107–109

Bradley, C. (1997). Psychological issues in clinical trial design. *Irish Journal of Psychology, 18*(1), 67–87.

Bradley, C. (1999). Patient preferences and clinical trial design and interpretation: Appreciation and critique of a paper by Feine, Awad and Lund. *Community Dentistry and Oral Epidemiology, 27,* 85–88.

Brewin, C. R., & Bradley, C. (1989). Patient preferences and randomised clinical trials. *British Medical Journal, 299,* 313–315.

Cooper, K. G., Grant, A. M., & Garratt, A. M. (1997). The impact of using a partially randomised patient preference design when evaluating alternative managements for heavy menstrual bleeding. *British Journal of Obstetrics and Gynaecology, 104,* 1367–1373.

Feine, J. S., Awad, M. A., & Lund, J. P. (1998). The impact of patient preference on the design and interpretation of clinical trials. *Community Dentistry and Oral Epidemiology, 26,* 70–74.

Henshaw, R. C., Naji, S. A., & Templeton, A. A. (1993). Comparison of medical abortion with surgical vacuum aspiration: Women's preferences and acceptability of treatment. *British Medical Journal, 307,* 714–717.

King, M., Nazareth, I., Lampe, F., Bower, P., Chandler, M., Morou, M., Sibbald, B., & Lai, R. (2005). Impact of participant and physician intervention preferences on randomized trials: A systematic review. *Journal of the American Medical Association, 293*(9), 1089–1099.

Torgerson, D. J., Klaber-Moffett, J., & Russell, I. T. (1996). Patient preferences in randomised trials: Threat or opportunity? *Journal of Health Services Research & Policy, 1*(4), 194–197.

## PARTICIPANTS

In the context of research, participants are individuals who are selected to participate in a research study or who have volunteered to participate in a research study. They are one of the major units of analysis in both qualitative and quantitative studies and are selected using either probability or nonprobability sampling techniques. Participants make major contributions to research in many disciplines. The manner in which their contributions are made is determined by the research design for the particular study and can include methodologies such as survey research, experiments, focus groups, and naturalistic observation. This entry focuses on the selection and protection of participants.

### Participant Selection and Recruitment

In research literature, the word *participants* is used often interchangeably with *informants, respondents, subjects,* or *interviewees*. However, there are some contextual distinctions between these terms. In survey research, participants are often referred to as respondents/interviewees. Respondents/interviewees provide information about themselves (e.g., opinions, preferences, values, ideas, behaviors, experiences) for data analysis by responding through self-administered surveys or interviews (e.g., telephone, face-to-face, email, fax). When experiments are being conducted, *subjects* are the preferred term for participants. Subjects are usually studied in order to gather data for the study. Finally, participants are considered to be informants when they are well versed in the social phenomenon of study (e.g., customs, native language) and are willing to speak about it. Overall, the distinction between a respondent and an informant is of particular importance because all respondents can talk about themselves; however, not all respondents can be good informants.

Participants can be identified/selected using two methods—probability sampling or nonprobability sampling. The method used for sample selection is dictated by the research questions. In probability sampling methodologies (e.g., simple random, systematic, stratified random, cluster), a random selection procedure is used to ensure that no systematic bias occurs in the selection process. This contrasts with nonprobability sampling methodologies (e.g., convenience, purposive, snowball, quota), where random selection is not used.

Once potential participants are identified, the next task involves approaching the individuals to try to elicit their cooperation. Depending on the purpose of the research, the unit of analysis may be either individuals or groups. For example, if

a researcher is examining student achievement based on test scores across two Primary 1 classes at School X, the unit of analysis is the student. In contrast, if the researcher is comparing classroom performance, the unit of analysis is the class because the average test score is being used.

The manner in which cooperation is elicited will vary according to the technique used to collect the data. However, the procedure generally involves greeting the participants, explaining the purpose of the study and how participants were selected, assuring participants that their responses will remain anonymous/confidential and that data will be aggregated prior to reporting, and asking them to participate. In other words, the actual recruitment involves giving participants sufficient information to enable them to give free and informed consent without being coerced. Participants must also be assured that they have the right to withdraw at any time.

In group situations (e.g., focus groups, classroom surveys) where potential participants are all in one location, cooperation may not be much of a problem. However, when random sampling is used and the unit of analysis is the individual, cooperation may be more difficult. For example, in telephone interviews, many potential participants may simply hang up the phone on telephone interviewers, and mail surveys may end up in the trash as soon as they are received. With face-to-face interviews, interviewer characteristics such as social skills and demeanor are partially important in trying to elicit cooperation. In general, the response rate for face-to-face interviews is higher compared to other methods, perhaps because participants may feel emoverlinerassed to close the door in the interviewer's face. When face-to-face interviews are being conducted, a letter is generally sent in advance to establish the legitimacy of the study and to schedule an appointment. However, when data are being collected nationwide (e.g., national census), public announcements on the radio/television and in local newspapers is the preferred route.

In research where participants have volunteered to participate (e.g., medical studies), cooperation is not an issue. However, other issues may take center stage. For example, all recruitment materials (e.g., press releases, radio/television/newspaper advertisements, flyers, posters) have to be approved by an institutional review board prior to publicity to ensure that eligibility criteria are clear; selection criteria are fair (i.e., some groups are not favored over others); advertising content is accurate/appropriate (e.g., if a study is comparing a new drug to a placebo, it must state that some participants will receive the placebo); benefits to participants are clear; any expected harm to participants (e.g., pain, distress) is outlined; and so on.

## Protection of Participants

In general, most research involving humans requires the approval of an institutional review board (IRB) to ensure that researchers are not violating the basic ethical rights of humans and/or causing harm to participants. The principles for ethical research are outlined in the "Belmont Report—Ethical Principles and Guidelines for the Protection of Human Subjects of Research." Of particular importance is research involving medical studies and research involving the use of children, the elderly, or persons who are mentally incompetent or have an impaired decision-making capacity. Research involving these types of participants is generally scrutinized in great detail by IRBs, so researchers have to provide very detailed protocol to the IRB in order to obtain approval. For the conduct of these types of research, consent may often need to be obtained from a third party. For example, in research involving children, parental/guardian consent must be obtained.

Although consent is a prerequisite for most types of research, there are some instances where it may not be desirable or possible. For example, in naturalistic observation, consent is not desirable because of the issue of reactivity, namely, where participants change their behavior when they are aware that they are being observed. In such research, it is incumbent that the researcher be mindful of the ethical principles as outlined in the Belmont Report and respect people's privacy and dignity.

*Nadini Persaud*

***See also*** Ethics in the Research Process; Interviewing; Naturalistic Observation; Nonprobability Sampling; Probability Sampling

**Further Readings**

Babbie, E. (2004). *The practice of social research* (10th ed.). Belmont, CA: Wadsworth/Thompson Learning.
Philliber, S. G., Schwab, M. R., & Sloss, G. S. (1980). *Social research: Guides to a decision-making process.* Itasca, IL: F. E. Peacock.

# Path Analysis

The use of path analysis to examine causal structures among continuous variables was pioneered by Sewall Wright and popularized in the social sciences through the work of Peter M. Blau and Otis D. Duncan, among others. There are several advantages to path analysis that account for its continuing popularity: (a) It provides a graphical representation of a set of algebraic relationships among variables that concisely and visually summarizes those relationships; (b) it allows researchers to not only examine the direct impact of a predictor on a dependent variable, but also see other types of relationships, including indirect and spurious relationships; (c) it indicates, at a glance, which predictors appear to have stronger, weaker, or no relationships with the dependent variable; (d) it allows researchers to decompose or split up the variance in a dependent variable into explained and unexplained, and also allows researchers to decompose the explained variance into variance explained by different variables; and (e) it allows researchers to decompose the correlation between a predictor and a dependent variable into direct, indirect, and spurious effects. Path analysis is used to describe systems of predictive or, more often, causal relationships involving three or more interrelated variables. Because path analysis is most often used in causal rather than purely predictive analysis, the language of causality is adopted here, but it should be borne in mind that causal relationships require more than path analysis for evidence; in particular, questions not only of association and spuriousness, but also of causal (and thus implicitly temporal) order must be considered.

In path analysis, more than one variable is typically treated as a dependent variable with respect to other variables in the model. Variables that affect other variables in the model, but are not affected by other variables in the model, are called *exogenous* variables, implying not so much that they are outside the model but that their explanation lies outside the model. A variable that is affected or predicted by at least one of the other variables in the model is considered an *endogenous* variable. An endogenous variable may be the last variable in the causal chain, or it may be an *intervening* variable, one that occurs between an exogenous variable and another endogenous variable. In practice, in any path analytical model, there will be at least one exogenous and one endogenous variable.

The simple patterns of direct, indirect, and spurious relationships are diagrammed in Figure 1. Diagram A in Figure 1 shows a simple relationship in which $X$ and $Y$ are both exogenous variables, and each has a direct effect on the endogenous variable $Z$. Diagram B in Figure 1 shows a spurious relationship, in which $X$ is an exogenous variable, $Y$ and $Z$ are endogenous variables, and $X$ has an effect on both $Y$ and $Z$. From Diagram B, we would expect that the zero-order correlation between $Y$ and $Z$ would be nonzero, but that the partial correlation between $Y$ and $Z$, controlling for $X$, would be zero. Diagram C in Figure 1 shows a causal chain in which $X$ is an exogenous variable and has a direct effect on $Y$ but not on $Z$, $Y$ and $Z$ are endogenous variables, and $Y$ has a direct effect on $Z$. In Diagram C, $X$ has an indirect effect on $Z$ through its effect on $Y$. We would therefore expect the zero-order correlation between $X$ and $Z$ to be nonzero, but the partial correlation between $X$ and $Z$, controlling for $Y$, to be zero. Diagram D in Figure 1 shows a mixture of direct and indirect effects, and incorporates all of the effects in the previous three diagrams. $X$ is an exogenous variable that has a direct effect on the endogenous variable $Y$, a direct effect on the endogenous variable $Z$, and an indirect effect (via $Y$) on the endogenous variable $Z$. $Y$ is an endogenous variable that is related to $Z$ in part through a direct effect, but also in part through a spurious effect because $X$ is a cause of both $Y$ and $Z$. In this diagram, we would expect the zero-order correlations among all the variables to be nonzero, the partial correlation between $X$ and $Z$ controlling for $Y$ to be nonzero but smaller than the zero-order correlation between $X$ and $Z$ (because part of the zero-order correlation would reflect the indirect effect of $X$ on $Z$ via $Y$), and the partial correlation between $Y$ and $Z$ controlling for

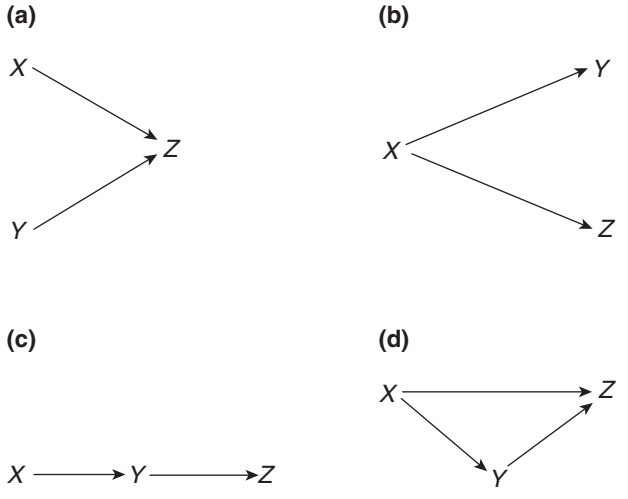



**Figure 1** Examples of Causal Relationships

*Notes:* (a) Direct effects of *X* and *Y* on *Z*. (b) Spurious effect of *X* on *Y* and *Z*. (c) Causal chain from *X* to *Y* to *Z*. (d) Direct plus indirect effects.

*X* to be nonzero but smaller than the zero-order correlation between *Y* and *Z* (because the zero-order correlation would reflect the spurious relationship between *Y* and *Z* resulting from their both being influenced by *X*).

Diagrams A and B in Figure 1 both represent models that can be described as *weakly ordered recursive* models. If we trace each arrow forward, we never come back to the variable where we started, but there are two variables (*X* and *Y* in Diagram A, *Y* and *Z* in Diagram B) between which there is not a clear causal ordering. In Diagram A, *X* and *Y* occur before *Z*, but it is not evident from the diagram that either occurs before the other. Similarly in Diagram B, *Y* and *Z* both occur after *X*, but there is no clear ordering between *Y* and *Z*. Diagrams C and D represent *strongly recursive* or *fully ordered recursive* models, in which it is again the case that if we trace each arrow forward, we never come back to the variable where we started, but it is also the case that *X* is the only exogenous variable in the model, and there is a clear ordering from left to right, with no ambiguity about which variable comes before or after any other variable. It is also possible to have *nonrecursive* models, in which it is possible to start from at least one variable, trace the arrows forward, and come back to the variable where you started. For example, in Diagram B in Figure 1, if we added two arrows, one from *Y* to *Z* and another from *Z* to *Y*, we would have a nonrecursive model, with a *nonrecursive loop* between *Y* and *Z*. Alternatively, in Diagram D in Figure 1, if we reversed the direction of the arrow from *X* to *Z* so it pointed instead from *Z* to *X*, we would have a nonrecursive loop involving all three of the variables (from *X* to *Y* to *Z* and back to *X* again).

## Decomposition of Variance in Path Analysis

In a model with three or more variables, *X*, *Y*, and *Z*, where *X* and *Y* are predictors of the dependent variable *Z* (corresponding to Diagram A in Figure 1), there may be (a) shared variance between *X* and *Y* that is not shared with *Z;* (b) shared variance between *X* and *Z* that is not shared with *Y;* (c) shared variance between *Y* and *Z* that is not shared with *X;* and (d) shared variance among all three variables, *X*, *Y*, and *Z*. The shared variance between *X* and *Y* that is not shared with *Z* represents the *zero-order covariance* between *X* and *Y*, and if expressed as a correlation instead of a covariance, it is the zero-order correlation between *X* and *Y*, the two exogenous variables in Diagram A of Figure 1. "Zero order" means that there are no (zero) other variables being statistically controlled in the computation of this correlation. The shared variance between *X* and *Z* that is not shared with *Y*, expressed as a correlation, is the *first-order partial correlation between X and Z, controlling for Y*, $r_{ZX \cdot Y}$. "First order" means that one variable (*Y*) is being controlled when calculating this measure of the association between *X* and *Z*. This partial correlation, if squared ($r^2_{ZX \cdot Y}$), may also be described as the variance in *Z uniquely* attributable to *X*. Similarly, the shared variance between *Y* and *Z* that is not shared with *X* is the first-order correlation between *Y* and *Z* controlling for *X*, $r_{ZY \cdot X}$. This partial correlation, if squared ($r^2_{ZY \cdot X}$), may also be described as the variance in *Z uniquely* attributable to *Y*. Finally, the variation shared among all three variables can be thought of as the *joint effect of X and Y on Z,* or the extent to which the effect of one exogenous variable on *Z* is *confounded* by (or not clearly distinguishable from) the effect of the other exogenous variable.

If *X* and *Y* are uncorrelated, the variation shared by all three variables is zero, and the partial

correlation of each exogenous variable with the dependent variable is the same as the zero-order correlation. In the more usual case when $X$ and $Y$ are correlated with each other, the variation shared by all three variables has traditionally been treated in two different ways. In analysis of variance and partial correlation analysis, it is attributed to neither $X$ nor $Y$, but treated as a separate source of explained variance in $Z$. In some standard analysis of variance output, the variance scenarios noted earlier in (b), (c), and (d) would each be presented as separate sources of explained variation in $Z$. In regression analysis, (d) is *partitioned* between $X$ and $Y$, and more of (d) will be attributed to whichever of $X$ and $Y$ has the higher overall (zero-order) correlation with $Z$. Unless $X$ and $Y$ provide a perfect prediction of $Z$, there will be variance in $Z$ not shared with $X$ and $Y$. The proportion of the total variance in $Z$ explained by $X$ and $Y$ is equal to the *squared multiple correlation of Z with X and Y, or* $R^2$. The variance in $Z$ that is not explained by $X$ and $Y$ (separately or in combination) is the *unexplained variance* (actually, the unexplained *proportion* of the variance), which can be written as $1 - R^2$, the square root of which, $\sqrt{1 - R^2}$, is sometimes explicitly included in a path model as the path coefficient for the "residual" effect, the presumed effect of variables not in the model that, if included, would explain the endogenous variable perfectly.

Path analysis, which is based on multiple linear regression analysis, like regression analysis partitions all of the explained variance in $Z$ among the predictors. According to this method of partitioning the variance, the proportion of the variance in $Z$ that is estimated to be shared *directly* (but not *uniquely*, because we are splitting up shared explained variance among the predictors) is equal to the product of the standardized regression coefficient and the zero-order correlation between the dependent variable and the predictor (e.g., $r_{ZX}b_{ZX}^*$ or $r_{ZY}b_{ZY}^*$). The sum of the products of all the standardized path coefficients and their associated zero-order correlations is known to be equal to the total proportion of the variance that is explained or $R^2$, that is, $\Sigma(b^*)(r) = R^2$. In this sense, the explained variance can be partitioned *completely*, but again not *uniquely*, among the predictors in the model, without specific reference to a separate shared effect.

## Methods of Estimating Path Coefficients

Methods of estimating path coefficients include (a) simultaneous estimation using structural equation modeling software, (b) separate estimation using ordinary least squares regression software when all variables in the model are (or can be treated as though they were) measured on an interval or ratio scale of measurement, and (c) separate estimation using logistic regression (and possibly ordinary least squares) when at least some of the variables in the model are dichotomous or categorical variables. In the earlier years of path analysis, the usual method of calculating path coefficients was to run separate regression analyses on each of the endogenous variables in the model. Scott Menard describes an approach to constructing path models using logistic regression analysis for categorical (dichotomous, nominal, or ordinal) variables that also relies on separate estimation, in which the categorical variables are treated as inherently categorical, rather than assuming that they are manifest indicators of latent (not directly measured, but estimated on the basis of observed indicator variables in the model) continuous variables.

When all of the variables in the model can be either (a) measured on a truly interval or ratio scale or (b) justifiably assumed to be categorical manifest indicators of continuous latent variables, the accepted practice has shifted to the use of software that simultaneously estimates all of the relationships in the model. The term *structural equation modeling* has come to refer less to the model itself than to the more sophisticated techniques for estimating simultaneous equation causal models. As described by David Heise, one major advantage of full-information maximum likelihood simultaneous equation methods for estimating path models is that they make more efficient use of valid theory, incorporating all such knowledge into the estimate of each system parameter, resulting in more precise estimates. The major disadvantage, correspondingly, is that they similarly incorporate any erroneous theory into the calculations, so if part of the model has been misspecified, all coefficient estimates may be affected instead of just a few. In general, Heise recommends the use of full-information methods in the latter stages of research, when there is a high level of confidence in the specification of the model. Estimation of

a single equation at a time may be more appropriate in the earlier stages of model testing, when the emphasis is less on precise estimation of coefficients for causal paths that are known with some certainty to belong in the model than on deciding *which* causal paths really belong in or are important to the model.

## Decomposition of Zero-Order Correlations in Path Analysis

Figure 2 presents the model from Diagram D of Figure 1 with coefficients assigned to each of the causal paths. One approach to calculating the coefficients for path analysis models is to first regress the last variable in the causal order ($Z$) on the prior two variables, then to regress each endogenous variable in turn, from last to second last, on the variables that precede it in the model. In this approach, the path coefficients (the numbers associated with each of the arrows) are identically the standardized regression coefficients produced by ordinary least squares regression analysis. In the model in Figure 2, $X$ has a direct effect (a path coefficient or standardized regression coefficient) on $Z$ of .200, and a direct effect on $Y$ of .100, and the direct effect of $Y$ on $Z$ is .300. The *indirect* effect of $X$ on $Z$ can be found by multiplying the path coefficients that connect $X$ to $Z$ via other variables. In this simple model, there is only one such path, and the indirect effect of $X$ on $Z$ is equal to $(b^*_{XY})(b^*_{YZ}) = (.100)(.300) = .030$. Alternatively, one can follow the common practice of using "$p$" in place of "$b^*$" to designate the path coefficients and rewrite this as $(p_{XY})(p_{YZ}) = .030$. The *total* effect of $X$ on $Z$ is equal to the *direct* effect plus the sum of all the *indirect* effects (of which there is only one in this example): $(p_{ZX}) + (p_{XYpYZ}) = (.200) + (.100)(.300) = .230$.

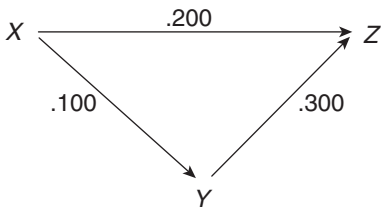



**Figure 2** Path Models and Decomposition of Correlations

*Note:* $r_{ZX} = .230$.

The fundamental theorem of path analysis states that the correlation between any pair of variables, such as $X$ and $Z$ or $Y$ and $Z$, can be expressed as the sum of all the nonredundant paths between the two variables. For a fully ordered recursive model, this is the sum of the direct plus indirect plus spurious causal paths between the two variables (i.e., the total effect plus the sum of the spurious effects) when the path coefficients are expressed as standardized regression coefficients. For a weakly ordered recursive model, this may also include paths involving correlations between exogenous variables, as described, for example, by Herbert B. Asher. It must also be the case that all nonzero paths are included in the model. When some paths are defined to be zero a priori, one test of how well the model works is to compare the *observed* correlation with the *implied* correlation (the latter calculated based on the path coefficients) between each of the predictors and the endogenous variables.

The total effect of $X$ on $Z$ in the present example will be equal to the zero-order correlation between $X$ and $Z$. Because $Y$ has no indirect effect on $Z$, the direct effect of $Y$ on $Z$ is the same as the total effect, .300, but there is also a spurious effect, equal to the product of the path from $X$ to $Y$ and the path from $X$ to $Z$: $(.100)(.200) = .020$, implying a zero-order correlation between $Y$ and $Z$ of $(.300) + (.020) = .320$. Finally, the only relationship between $X$ and $Y$ is the direct effect of $X$ on $Y$, and the zero-order correlation in this instance should be equal to the path coefficient for that direct effect, .100. Because the path coefficients in Figure 2 are standardized regression coefficients, it is reasonable to conclude that the impact (measured in terms of the total effect) of $Y$ on $Z$ is stronger than the impact of $X$ on $Z$, even when we include direct as well as indirect effects.

Now consider some possible changes in the model. If the path coefficient between $X$ and $Z$ were .280 instead of .200, then the total effect of $X$ on $Z$ would be $(.280) + (.030) = .310$, and the total effect of $X$ on $Z$ (.310) would be slightly larger than the total effect of $Y$ on $Z$ (.300), even though the zero-order correlation between $Y$ and

$Z$ (.320, equal to the direct plus spurious effect) is larger than the zero-order correlation between $X$ and $Z$ (.310, equal to the direct plus indirect effect). Second, if the direct effect of $X$ on $Z$ were zero instead of .200, $X$ would still have the indirect effect, .030, on $Z$, and because it had only an indirect effect, the indirect effect would be equal to the total effect (and to the zero-order correlation) between $Z$ and $X$. Third, if the direct effect of $X$ on $Y$ were zero, then the direct effect of $X$ on $Z$ would be equal to the total effect of $X$ on $Z$, .200, and moreover, with $X$ and $Y$ uncorrelated, the path coefficients would be equal not only to the standardized regression coefficients but also to the zero-order correlations for the relationship between $X$ and $Z$ and the relationship between $Y$ and $Z$. All of these are relatively simple examples. In general, it will not usually be the case that the zero-order correlation coefficient will be equal to the standardized regression coefficient, that predictors will be completely uncorrelated, or that total effects will be equal to either the direct or the indirect effect.

*Scott Menard*

***See also*** Correlation; Multiple Regression; Structural Equation Modeling

**Further Readings**

Asher, H. B. (1983). *Causal modeling*. Beverly Hills, CA: Sage.

Blau, P. M., & Duncan, O. D. (1967). *The American occupational structure*. New York: Wiley.

Heise, D. R. (1975). *Causal analysis*. New York: Wiley.

Menard, S. (2010). *Logistic regression: Introductory to advanced concepts and applications*. Thousand Oaks, CA: Sage.

Tatsuoka, M. M. (1971). *Multivariate analysis: Techniques for educational and psychological research*. New York: Wiley.

Wright, S. (1918). On the nature of size factors. *Genetics, 3,* 367–374.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research, 20,* 557–585.

Wright, S. (1934). The method of path coefficients. *Annals of Mathematical Statistics, 5,* 161–215.

Wright, S. (1971). Path coefficients and path regressions: Alternative or complementary concepts. *Biometrics, 16,* 189–202.

# Pearson Product-Moment Correlation Coefficient

In the late 19th century, Sir Francis Galton was measuring many different biological and sociological variables and describing their distributions using the extant methods for single variables. However, there was no way to quantify the degree of relationship between two variables. The Pearson product-moment correlation coefficient (hereafter referred to as "coefficient") was created by Karl Pearson in 1896 to address this need. The coefficient is one of the most frequently employed statistical methods in the social and behavioral sciences and is frequently used in theory testing, instrument validation, reliability estimation, and many other descriptive and inferential procedures.

This entry begins by defining the coefficient and describing its statistical properties. Next, this entry discusses descriptive and inferential procedures. It closes with a discussion of the coefficient's limitations and precautions for interpreting.

## Definition

The population coefficient, which is typically abbreviated as $\rho$, or the Greek letter rho, is an index of the degree and direction of linear association between two continuous variables. These variables are usually denoted as $X$ (commonly labeled as a predictor variable) and $Y$ (commonly labeled as an outcome variable). Note, however, that the letters are arbitrary; the roles of the variables implied by these labels are irrelevant because the coefficient is a symmetric measure and takes on the same value no matter how two variables are declared.

The population value is estimated by calculating a sample coefficient, which is denoted by $\hat{\rho}$, "rho hat," or $r$. The sample coefficient is a descriptive statistic; however, inferential methods can be applied to estimate the population value of the coefficient.

## Statistical Properties of the Coefficient

The coefficient can take on values between –1 and 1 (i.e., $-1 \leq \rho \leq 1$). The absolute value of the coefficient denotes the strength of the linear association

with $|\rho| = 1$ indicating a perfect linear association and $\rho = 0$ indicating no linear association. The sign of the coefficient indicates the direction of the linear association. When high scores on $X$ correspond to high scores on $Y$ and low scores on $X$ correspond to low scores on $Y$, the association is positive. When high scores on $X$ correspond to low scores on $Y$, and vice versa, the association is negative.

Bivariate scatterplots are often used to visually inspect the degree of linear association between two variables. Figure 1 demonstrates how an $(X,Y)$ scatterplot might look when the population correlation between $X$ and $Y$ is negligible (e.g., $\rho = .02$). Note that there is no observable pattern in the dots; they appear to be randomly spaced on the plot.

Figure 2 demonstrates how an $(X,Y)$ scatterplot might look when $\rho = .90$. There is an obvious pattern in the paired score points. That is, increases in $X$ are associated with linear increases in $Y$, at least on average.

Figure 3 demonstrates how an $(X,Y)$ scatterplot might look when $\rho = -.30$. In this plot, increases in $X$ are associated with linear decreases in $Y$, at least on average. Due to the lower absolute value of the coefficient in this case, the pattern is not as easily identifiable.

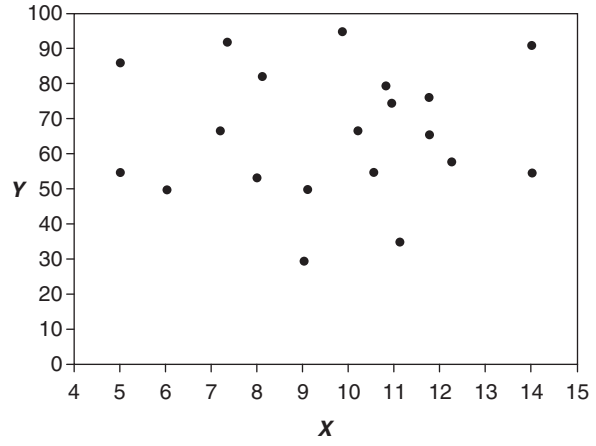



**Figure 1** No Correlation

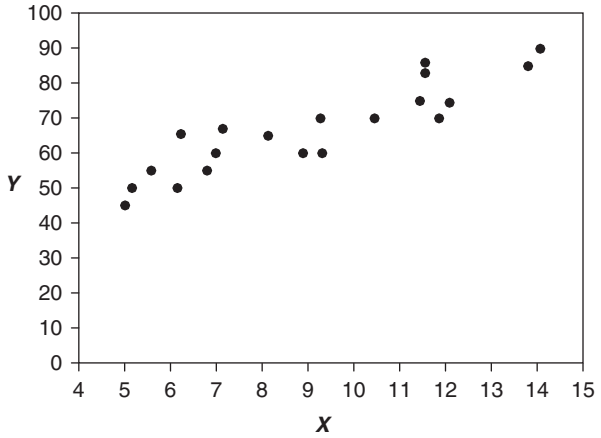



**Figure 2** Strong Positive Correlation

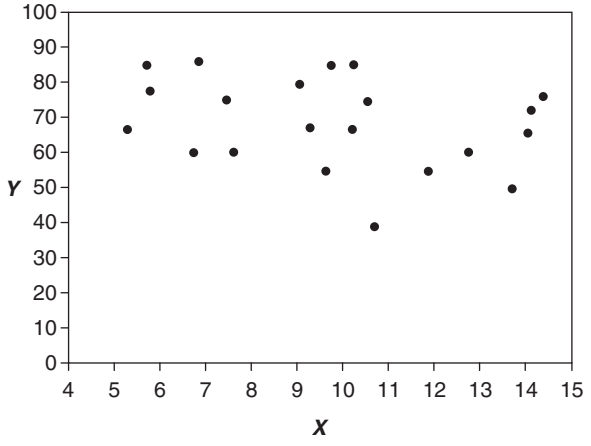



**Figure 3** Moderate Negative Correlation

### *Shared Variance*

In addition to the strength and direction of linear association, the coefficient also provides a measure of the shared variance between $X$ and $Y$, which is calculated by squaring the coefficient. The resulting value, $r^2$, is also referred to as the *coefficient of determination*. Specifically, $r^2$ indicates the percentage of variance in one variable that is attributable to the variance in the other variable. For example, if the correlation between $X$ and $Y$ is $\rho = .90$, then $r^2 = .81$; therefore, 81% of the variance in $X$ scores can be predicted from the variance in $Y$ scores.

## Descriptive Procedures for *r*

### *Suitability of* r

The appropriateness of computing $r$ is based on the assumption that $X$ and $Y$ are both continuous variables (i.e., $X$ and $Y$ are measured on either

interval or ratio scales) and that the relationship between $X$ and $Y$ can be best represented by a linear function. It is possible for two variables to have a linear or nonlinear form of relationship; however, the coefficient only quantifies the direction and degree of linear relationship between two variables. If the association between $X$ and $Y$ is curvilinear, the coefficient can still be computed but does not provide the optimal representation of the relationship. In the case of a nonlinear monotonic relationship, Spearman's correlation coefficient is a better measure of the degree of monotonic association between $X$ and $Y$. Incidentally, Spearman's correlation coefficient is mathematically the same as computing Pearson's correlation coefficient on the ranks of the data.

### *Calculating the Value of* r

The coefficient provides an index of the consistency with which scores on $X$ are linearly associated with scores on $Y$. If scores on $X$ are linearly associated with scores on $Y$, then the location of an individual's score in the $X$ distribution (i.e., his or her standardized $z$ score on $X$) will be similar to (or opposite of, in the case of an inverse relationship) the location of the individual's score in the $Y$ distribution (i.e., his or her standardized $z$ score on $Y$). Therefore, the coefficient is calculated by finding the average cross-product of the $z$ scores of $X$ and $Y$:

$$\rho = \frac{\sum_{i=1}^{N} Z_X Z_Y}{N} \text{ in the population and}$$

$$r = \frac{\sum_{i=1}^{n} z_X z_Y}{n-1} \text{ in the sample,}$$

where $Z_X$, $Z_Y$, $z_x$, and $z_Y$ are the $z$ scores for $X$ and $Y$ in the population and sample, respectively, and $N$ and $n$ are the population and sample sizes, respectively. This fomula can also be re-expressed as the ratio of the covariance of $X$ and $Y$ to the product of the standard deviations of $X$ and $Y$; in other words, it can be expressed as the ratio of how much $X$ and $Y$ vary together compared to how much $X$ and $Y$ vary separately:

$$\rho = \frac{Cov(X,Y)}{SD(X)SD(Y)},$$

where $SD$ denotes standard deviation.

## Inferential Procedures for $\rho$ Using $r$

### *Calculating Confidence Intervals for* $\rho$ *Using* r

It is possible to calculate a confidence interval for $\rho$ using the value of the sample coefficient $r$. However, to achieve proper confidence coverage, the value of $r$ has to be transformed using what is known as Fisher's $z$ transformation, here denoted $z'$,

$$z' = \frac{1}{2}\left[\ln\left(\frac{1+r}{1-r}\right)\right]$$

where ln is the natural logarithm. Because the asymptotic sampling distribution for Fisher's $z$ values is known to be the normal distribution with standard error

$$SE = \frac{1}{\sqrt{n-3}},$$

a confidence interval for $\rho$ on the transformed scale can be constructed in the prototypical manner via

$$z' \pm \text{TLV} \cdot \frac{1}{\sqrt{n-3}},$$

where TLV is the table-look-up value associated with the desired confidence level of the interval (e.g., $\text{TLV}_{95\%} = 1.96$, $\text{TLV}_{90\%} = 1.65$). Once the confidence interval for the value of $\rho$ on Fisher's $z$ scale has been constructed, these bounds of the confidence interval must be transformed back into bounds for the confidence interval for $\rho$ by using

$$\rho_{Lower} = \frac{\exp(2 \cdot z'_{Lower}) - 1}{\exp(2 \cdot z'_{Lower}) + 1} \text{ and}$$

$$\rho_{Upper} = \frac{\exp(2 \cdot z'_{Upper}) - 1}{\exp(2 \cdot z'_{Upper}) + 1},$$

where $z'_{Lower}$ and $z'_{Upper}$ are the confidence bounds of the confidence interval for $\rho$ on Fisher's $z$ scale.

### *Testing $H_0 : \rho = 0$*

To test the null hypothesis that the population correlation between $X$ and $Y$ is 0 (i.e., to test $H_0 : \rho = 0$ vs. $H_1 : \rho \neq 0$), one can use the $t$-distribution by calculating a $t$ score for $r$ using

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}},$$

where $r$ is the calculated correlation coefficient,

$$\frac{\sqrt{1-r^2}}{\sqrt{n-2}}$$

is the standard error of $r$, and $n$ is the sample size. One then compares this calculated $t$ value to a critical $t$ value with $df = n - 2$ at a desired alpha level. In addition, most statistical textbooks provide lookup tables in which to locate critical values for $r$ without having to convert $r$ to a $t$ score.

### *Testing a Nonzero Null Hypothesis*

Researchers may want to test a null hypothesis other than $H_0$: $\rho = 0$ to determine if a calculated coefficient is significantly different from some other value, for example, .20 (i.e., they may want to test $H_0 : \rho = .20$ vs. $H_1 : \rho - .20$). To do this, the researcher must convert both the calculated and critical values of $r$ to a $z$ score using Fisher's $z$ transformation and then use the prototypical fomula for a test statistic to conduct the test:

$$\text{Test Statistic} = \frac{z'_{(\text{calculated})} - z'_{(\text{hypothesized})}}{\frac{1}{\sqrt{n-3}}},$$

where $z'_{(\text{calculated})}$ is the $z$-transformed $r$ value calculated from sample data, $z'_{(\text{hypothesized})}$ is the $z$-transformed value hypothesized under the null, and the expression in the denominator is the standard error of $r$ on Fisher's $z$ scale.

### *Testing for a Difference Between Two Independent Correlations*

One can also test to determine if two independent correlations are significantly different by calculating a $z$ score of the difference between the $z$ transformed correlations in an analogous manner:

$$\text{Test Statistic} = \frac{z'_{\text{Sample 1}} - z'_{\text{Sample 2}}}{\frac{1}{\sqrt{(n_1-3)+(n_2-3)}}},$$

where $z'_{\text{Sample 1}}$ and $z'_{\text{Sample 2}}$ are the $z$-transformed correlation coefficients for Sample 1 and Sample 2, respectively, and $n_1$ and $n_2$ are the sample sizes for Samples 1 and 2, respectively.

## Assumptions for Inferential Procedures

The correctness of the inferential procedures in the preceding section is based on the following set of assumptions that have to hold in addition to the two variables being continuous and the relationship being linear as stated earlier.

1. *The scores of* X *and* Y *are sampled from a bivariate normal distribution.* Bivariate normality refers to the condition in which all individuals' $Y$ scores are normally distributed in the population across all $X$ scores. That is, if one examined the $Y$ scores of all individuals with an $X$ score of, for example, 50, one should find that, assuming bivariate normality holds, these individuals' $Y$ scores are normally distributed. Essentially, bivariate normality is obtained when both variables are normally distributed. When one or both variables are skewed, bivariate normality is often violated, possibly adversely affecting the inferential properties of the coefficient.

2. *Homoscedasticity.* Homoscedasticity refers to the condition in which, for all given $X$ values, the amount of variation in $Y$ values remains relatively constant. Homoscedasticity would be violated if, for example, at the lower end of the $X$ scale, there was very little variation in $Y$ scores, but at the upper end of the $X$ scale, there was a large amount of variation in the $Y$ scores.

3. *The sample of scores has been randomly selected from the population of interest.* If the sample is not randomly selected, it is uncertain to what degree the calculated coefficient represents the population value, because the calculated value may be systematically biased upward or downward.

### Limitations of the Coefficient and Precautions Regarding Its Interpretation

#### *Outliers*

The value of the coefficient is quite susceptible to outliers. The presence of even one outlier can spuriously attenuate or inflate the true value of the correlation. Therefore, it is necessary to attempt detection of outliers when wishing to draw conclusions based on a calculated coefficient.

#### *Restriction of Range*

In addition to outliers, restriction of range can have unwanted effects on the value of the coefficient. That is, when only a portion of the entire range of values of a variable is included in the sample (i.e., if the sample data are not representative of the population data), the calculated coefficient may be an inaccurate estimate of the population value. Therefore, one must be careful to consider the range of possible values of the $X$ and $Y$ variables and to take into account to what degree that range is represented in one's current sample.

#### *Correlation and Causation*

It is important to emphasize that although a correlation coefficient may suggest that two variables are associated, it does not prove any causal relationship between the two variables. Researchers must be careful about conclusions drawn from correlation coefficients; even when $r = 1.00$, one cannot be sure that one variable *causes* another, only that one variable is *related* to another.

*Michael J. Walk and André A. Rupp*

***See also*** Coefficients of Correlation, Alienation, and Determination; Correlation; Cross-Sectional Design; Homoscedasticity; Regression to the Mean; Scatterplot

#### Further Readings

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* New York: CBS College Publishing.

Miles, J., & Banyard, P. (2007). *Understanding and using statistics in psychology: A practical introduction.* London: Sage.

Sheskin, D. J. (2000). *Handbook of parametric and nonparametric statistical procedures* (2nd ed.). New York: Chapman & Hall/CRC.

Thorndike, R. M. (1978). *Correlational procedures for research.* New York: Gardner.

Wilcox, R. R. (2001). Modern insights about Pearson's correlation and least square regression. *International Journal of Selection and Assessment, 9,* 195–205.

Wilcox, R. R., & Muska, J. (2001). Inferences about correlations when there is heteroscedasticity. *British Journal of Mathematical and Statistical Psychology, 54,* 39–47.

## Percentile Rank

There is no standard definition of a percentile, but three common definitions yield similar, if not identical, results when working with a large sample size. Some define a percentile to be the smallest value that a specified percentage of the observations is less than. Another definition is that a percentile is the smallest value that a specified percentage of the observations is less than or equal to, and this is commonly what is referred to as the percentile rank. Conveniently, a third definition that uses both of the previous two exists and handles small data sets consistently. This third method is recommended by the National Institute of Standards and Technology (NIST) for identifying percentiles and is used here.

### Calculation

#### *Notation*

| | |
|---|---|
| $Y(p)$ | the $p$th percentile |
| $n$ | the sample size, or the number of observations |
| $r$ | the rank, or position in an ordered list, of an observation |
| $Y_{(r)}$ | the observation with rank $r$ in the sample (e.g., $Y_{(1)}$ is the smallest observation; $Y_{(n)}$ is the largest) |
| $p$ | the percentage of interest |

#### *Fomula*

First, sort all $n$ raw scores in ascending order. Next, find the rank ($r$), or position in the

ordered listing of scores, of the percentile of interest ($p$):

$$r = \frac{p}{100}(n+1)$$

To find the score representing that percentile, $Y(p)$,

$$Y(p) = \begin{cases} Y_{(1)} & k = 0 \\ Y_{(n)} & k = n \\ Y_{(k)} + d\left(Y_{(k+1)} - Y_{(k)}\right) & 0 < k < n, \end{cases}$$

where $k$ is the integer portion of $r$, and $d$ is the decimal portion of $r$, such that $k + d = r$.

### *Example*

The raw scores of a test administered to seven people are listed below, along with their respective ranks. You are interested in finding the 80th percentile.

| *Score* | *Rank* |
|---|---|
| 85 | 4 |
| 77 | 1 |
| 88 | 5 |
| 91 | 7 |
| 90 | 6 |
| 81 | 2 |
| 83 | 3 |

The rank, $r$, of the 80th percentile is

$$r = \frac{p}{100}(n+1) = \left(\frac{80}{100}\right)(7+1) = 6.4.$$

Using the definitions above, $k = 6$ and $d = 0.4$, giving $k + d = r$. The 80th percentile can now be calculated as follows:

$$\begin{aligned} Y(80) &= Y_{(k)} + d\left(Y_{(k+1)} - Y_{(k)}\right) \\ &= Y_{(6)} + d\left(Y_{(7)} - Y_{(6)}\right) = 90 + .4(91 - 90) \\ &= 90.4. \end{aligned}$$

That is, an examinee who received a raw score of 90.4 on this exam performed as well as or better than 80% of the other test takers. For a sample size of only seven people, one can see that this is neither very informative nor useful. However, when dealing with larger samples, percentiles can be very helpful in describing where one stands in relation to others.

## The Sample Median

The sample median is a useful statistic in that, unlike the mean and mode of a set of observations, it is unaffected by outliers and skewness in the data because it depends only on the values in the very middle of all the observations. The median is often used to report statistics whose distributions are widely varied or skewed, such as annual household income for a community or the ages of students enrolled in a program.

By definition, the median is the halfway point of a data set: Exactly half of the observations are less than it, and therefore, exactly half are greater than it. Thus, the median is also the 50th percentile of a set of data.

### *Example*

One can see that, if $r$ is an integer, no interpolation is necessary, and the percentile is the value with rank $r$. An example illustrates this point by finding the 50th percentile, or median, of the seven scores above.

$$r = \frac{p}{100}(n+1) = \left(\frac{50}{100}\right)(7+1) = 4.0$$

Now, $k = 4$ and $d = 0.0$. To find the median,

$$\begin{aligned} Y(50) &= Y_{(k)} + d\left(Y_{(k+1)} - Y_{(k)}\right) \\ &= Y_{(4)} + d\left(Y_{(5)} - Y_{(4)}\right) = 85 + 0(88 - 85) \\ &= 85. \end{aligned}$$

Thus, exactly half of the seven test scores are below 85, and the other half are greater than 85.

## Other Special Cases

In addition to the median, other special cases of percentiles are often used to classify students, schools, or others. Besides the median, the most commonly used percentiles are the first and third quartiles. As their name suggests, quartiles divide a data set into fourths. That is, the first quartile

(denoted $Q_1$) signifies the value with one quarter of the data less than or equal to it; the third quartile ($Q_3$) is the value with three quarters of the data less than or equal to it. Therefore, to find the first quartile of a data set, one need only find the 25th percentile. Similarly, finding the third quartile is the same as finding the 75th percentile. The median is the second quartile ($Q_2$).

Another, less common, special case of percentiles are deciles, which divide the data into tenths. They can likewise be found using the fomula above. The first example, finding the 80th percentile, was also an example of finding the eighth decile of the exams.

## Uses of Percentiles

Many tests and instruments use percentile ranks for the gradation and reporting of scores to examinees. A percentile rank score is neither a raw score, like in the examples above, nor a proportion or percent correct related to an examinee's performance on an instrument. Instead, the percentile rank is a measure of an individual's performance level relative to a reference (or comparison) group. The reference group is designated with consideration to the test's purpose. For example, when administering an educational standards test, the reference group could be the general location in which a school is located, an entire state, or an entire nation. The score that is reported is the $p$th percentile, indicating how well the examinee performed on the test relative to the rest of the reference group.

National and local standardized tests often include a percentile rank in addition to a raw score when reporting an individual's examination results. As illustrated below, one use of these scores is to see where an individual stands among his or her peers who also took the exam. Another use allows others, such as college admissions boards, to compare results from the current examination with students who took other versions of the test, or who took it in earlier or later years.

Another use of percentiles is for the classification of data. As previously stated, a person with a percentile rank of $p$ splits the data into two groups, where $p$% of the reference group received the same or lower score while the remaining $(100 - p)$% of the group received higher scores. The percentile rank can also be used to mark a minimum level of proficiency, or a cut score. If a test is designed to single out the top $p$% of individuals possessing a level of a trait, those examinees who receive a percentile rank of $(100 - p)$ or greater should be investigated for inclusion.

A test that rank orders an examinee based on percentile ranks is said to be norm-referenced, because the examinee is only compared to others, and the examinee's raw score is not compared against some pre-set criterion. Percentile ranks are often used in norm-referenced settings, as when, for example, determining which students are at the head of their class, or deciding who the top-performing employees are when awarding bonuses or raises. Contrarily, if a school required the same minimum score on a test from year to year for admission, or if a company awarded bonuses only to salespeople who brokered a given number, this would be a criterion-referenced situation.

## Test Equating

Percentile rankings are not only used when reporting results of standardized tests. In the field of test equating, one is interested in equating the set of scores for one instrument to those scores found on another instrument. There are many versions of equating found in classical test theory (CTT), such as mean, linear, and equipercentile test equating. In CTT, the beauty of equipercentile equating is that it is a nonlinear method.

Consider the situation of having two different tests, Test A and Test B, that measure the same constructs but consist of different questions. Table 1 shows an example of what the two test forms A and B might look like. For both tests, the raw score is reported along with the proportion of examinees who received that score and the cumulative proportion of all scores at that point or lower. The cumulative proportion of examinees for each score is the same as the percentile ranking of that score. One can immediately see that the two tests are not equivalent: Test A contains 10 questions and Test B contains 15 questions. This fact alone will give the possibility of a higher raw score on Test B, simply due to the larger number of questions.

In order to be able to compare performance on Test A to performance on Test B, one can compare

**Table 1** Raw Scores for Test A and Test B

| *Test A* | | | *Test B* | | |
|---|---|---|---|---|---|
| *Raw Score* | *Prob(Raw Score)* | *Cumulative* | *Raw Score* | *Prob(Raw Score)* | *Cumulative* |
| 0 | 0.05 | 0.05 | 0 | 0.05 | 0.05 |
| 1 | 0.20 | 0.25 | 1 | 0.05 | 0.10 |
| 2 | 0.10 | 0.35 | 2 | 0.05 | 0.15 |
| 3 | 0.15 | 0.50 | 3 | 0.05 | 0.20 |
| 4 | 0.05 | 0.55 | 4 | 0.05 | 0.25 |
| 5 | 0.05 | 0.60 | 5 | 0.10 | 0.35 |
| 6 | 0.05 | 0.65 | 6 | 0.10 | 0.45 |
| 7 | 0.05 | 0.70 | 7 | 0.10 | 0.55 |
| 8 | 0.05 | 0.75 | 8 | 0.10 | 0.65 |
| 9 | 0.05 | 0.80 | 9 | 0.05 | 0.70 |
| 10 | 0.20 | 1.00 | 10 | 0.05 | 0.75 |
| | | | 11 | 0.05 | 0.80 |
| | | | 12 | 0.05 | 0.85 |
| | | | 13 | 0.05 | 0.90 |
| | | | 14 | 0.05 | 0.95 |
| | | | 15 | 0.05 | 1.00 |

percentiles instead of raw scores. The most direct way of finding equivalent scores for these two test forms is to compare the raw scores from each test that have equivalent percentiles.

### *Example*

Given the two tests, A and B, as outlined in Table 1, compare the 65th percentiles for the two forms. On Test A, a raw score of six questions correct is the 65th percentile; on Test B, a raw score of eight questions correct is the 65th percentile. Although the raw score on Test B is higher than that for Test A, the scores are considered to be equivalent because they are the same percentile; an examinee with a raw score of 6 on Test A and an examinee with a raw score of 8 on Test B have done equally well, relative to all other participants. Likewise, if an examinee were to correctly answer six questions on Test A, then he or she would likely answer approximately eight questions correctly if the examinee were to take Test B, and vice versa. The method of equating tests via matching percentiles is called equipercentile equating.

In order to illustrate the nonlinearity of the equipercentile equating method, one can pair raw scores by matching the percentile ranks from the tests being equated and then plot these matched pairs for a graphical representation, as in Figure 1. In the case of Test A and Test B in our example, one can pair raw scores from the 5th, 25th, 35th, 55th, 65th, 70th, 75th, and 99th percentiles, which are the only percentiles exactly matched on both tests. This graph can be used to estimate the expected score on an alternate test form when only one form had been administered, and one can interpolate equal raw scores on two exams. For example, if an examinee gets a raw score of 5 on Test A, we can see that that is the equivalent of a raw score of around 7.5 on Test B.

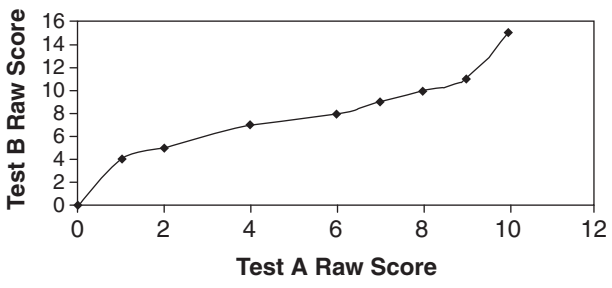



**Figure 1** Test A and Test B Equipercentile Equating

One major drawback to equipercentile equating, however, is that estimation for raw scores higher than or lower than those actually observed is not possible because all percentiles are accounted for in the data actually at hand. While one is often not interested in the properties of a test—like discrimination—for those obtaining low scores, it does pose problems for tests where high scores are of interest.

*Megan Lutz and Mike McGill*

***See also*** Descriptive Statistics; Mean; Standard Deviation

**Further Readings**

Anderson, D. R., Sweeney, D. J., & Williams, T. A. (1991). *Introduction to statistics* (2nd ed.). New York: West.

Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 9–49). New York: Academic Press.

David, H. A., & Nagaraja, H. N. (2003). *Order statistics* (3rd ed.). Hoboken, NJ: Wiley.

Devore, J. L. (2007). *Probability and statistics for engineering and the sciences* (7th ed.). Belmont, CA: Thomson Higher Education.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (pp. 36–48). New York: Springer.

National Institute of Standards and Technology & Semiconductor Manufacturing Technology. (2006). *NIST/SEMATECH e-handbook of statistical methods.* Retrieved from http://www.itl.nist.gov/div898/handbook/prc/section2/prc252.htm

Wang, J. (1995). Toward clarification of confusion in the concept of percentile. *Education, 115*(4), 538–540.

# Pie Chart

A pie chart is a way of displaying data in which a circle is divided into segments (or "slices") that reflect the relative magnitude or frequency of the categories. For example, the world's population is divided among the continents as in Table 1. In a pie chart based on these data (Figure 1), the segment for Africa would constitute 13.72% of the total world population, that for Asia 60.63%, and so forth.

Pie charts are most often used to display categorical (i.e., nominal) data, and less often ranked or ordinal data. They are never used to show continuous data, for which line charts are the obvious choice. After a brief history of the pie chart, this entry describes variations of and problems with pie charts and then enumerates guidelines for making better pie charts.

## A Brief History

The pie chart was probably first used by William Playfair, in a book published in 1801 called *The Statistical Breviary.* In it, he used circles of different sizes to represent the areas of the various countries in Europe and how they changed over time.

**Table 1** Population of the World's Continents

| *Continent* | *Population (in 000,000s)* | *Percent of Total* |
|---|---|---|
| Africa | 878 | 13.72 |
| Asia | 3,879 | 60.63 |
| Europe | 727 | 11.36 |
| North America | 502 | 7.85 |
| Oceania | 32 | 0.50 |
| South America | 380 | 5.94 |
| *Total* | 6,398 | 100 |

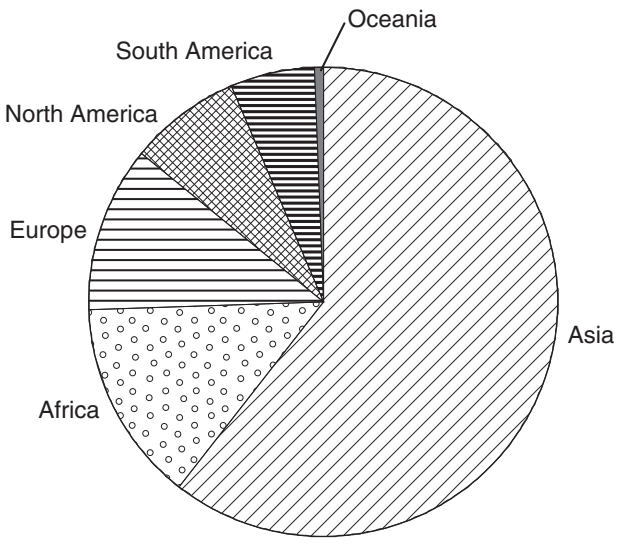



**Figure 1** Pie Chart Based on the Data in Table 1

The circles were colored to indicate whether the country was a maritime nation (green) or not (red), and some were divided into slices to reflect the ethnic backgrounds of the residents. In a previous book, Playfair had used overline charts and line charts for the first time. However, possibly due to his reputation as a shady businessman, pie charts were not adopted in the United Kingdom for nearly 100 years, although they were better received in Europe.

## Variations

In an *exploded pie chart,* one segment is separated from the rest. This can be done either to highlight that segment as deserving of special attention, or because that slice is so small that it would otherwise be overlooked. In a *polar area diagram,* all of the segments have the same angle, and differ from each other in terms of how far each slice extends from the center. Although this type of chart is often attributed to Florence Nightingale (and, indeed, is sometimes referred to as a Nightingale rose diagram), it was likely first used in 1843 by Léon Lalanne.

## Problems With Pie Charts

Despite their ubiquity in newspapers, magazines, and business presentations, pie charts are rarely used in the sciences. They are pleasing to look at, but it is difficult to draw conclusions about the numbers that they represent. One major problem is the interpretation of the area of the individual segments. Viewers can easily discriminate among lines of different lengths, as in overline charts and histograms. They are less accurate, though, in discerning differences among areas and angles. Doubling the length of a line makes it look twice as long, but doubling an area makes it appear 70% larger. If the slices of a pie chart differ from each other only slightly, people have much more difficulty rank-ordering them than if the data were presented as overlines. Viewers also have difficulty interpreting angles; they tend to underestimate acute angles and overestimate obtuse ones. Furthermore, the same angles are seen as different if they are oriented differently (e.g., the slice extending vertically versus horizontally from the center).

Some chart users attempt to avoid this problem by placing the percentage or count either inside the slice or just outside it. However, this does violence to the purpose of a graph, which is to allow the viewer to quickly see relationships among categories. If numbers are required in order for the viewer to do this, it would be better to use a table rather than a graph.

A second problem is that it is nearly impossible to compare the size of slices across two or more pie charts. Unless the slices both begin at the 12-o'clock position, comparing them imposes a heavy cognitive demand. The viewer is required to mentally rotate the slice in one pie until its left margin is at the same angle as that of the comparable slice in the other pie, while at the same time preserving the internal angle of the slice, and then judging whether the right margins are at the same angle. This is not just difficult—it is nearly impossible to do accurately.

## Rules for Making Better Pie Charts

If a pie chart is used, there are a number of guidelines for making them easier to interpret.

1. The number of categories should be relatively small. If there are only two categories (e.g., male/female, responders/nonresponders), a chart of any type is superfluous, and the information would be better presented within the text. When there are more than four to six categories, the segments become too small to easily differentiate.
2. Values that differ considerably from one another are easier to differentiate than values that are nearly identical. If some segments have similar areas, a different type of graph should be considered.
3. The names of the segments should be placed near the segments themselves, not in a legend box alongside or beneath the chart. It is cognitively much more demanding for the viewer to look at a segment, remember its color or shading, and then search for that attribute in a list of the categories. However, many widely available computer programs that draw graphs cannot do this easily and place the legend in a separate box.

4. Unless the data are ordinal, the segments should be arranged so that they are in descending order of magnitude. The first should start at the 12-o'clock position, and the slices should be arranged in a clockwise direction.
5. Three-dimensional (perspective) pie charts should never be used. Although they are often found in nonscientific publications and presentations, they are extremely misleading. Tilting the angle distorts the perceived relationships among the segments, and the greater the degree of perspective, the greater the distortion. Especially when the segments are in different colors and one or more slices are exploded, the viewer often focuses more on the design than on the content.

*David L. Streiner*

***See also*** Graphical Display of Data; Histogram; Line Graph

**Further Readings**

Cleveland, W. S. (1985). *The elements of graphing data.* Pacific Grove, CA: Wadsworth.
Spence, I. (2005). No humble pie: The origins and usage of a statistical chart. *Journal of Educational and Behavioral Statistics, 30,* 353–368.
Tufte, E. (2001). *The visual display of quantitative information.* Chesire, CT: Graphics Press.

# Pilot Study

In research, a pilot study refers to either a trial run of the major research study or a pretest of a particular research instrument or procedure. Ideally, such studies should be conducted using participants who closely resemble the targeted study population. Pilot studies are particularly valuable in situations where little is known about the research topic, or when executing unprecedented research instruments. The major objective of a pilot study is to discover problems prior to the main study so that the researcher can take corrective action to improve the research process, and thus the likelihood of success of the main study. This entry discusses the importance of pilot studies, the procedures and evaluation of pilot studies, and problems associated with them.

## Importance

All researchers seek to obtain reliable and valid data to answer their research questions or hypotheses. However, although researchers generally try to be quite meticulous, measurement error can still easily occur as a result of problems with questionnaire design, improperly trained interviewers, and so on. Therefore, pilot studies should be a normal component of good research design. Such studies can save researchers both time and money because logistical problems and other design deficiencies can be identified prior to the real study, and corrections and adjustments can be made before the main study is executed.

In some cases, more than one pilot study may be necessary. For example, the first pilot study may be an expert review or focus group to help determine and/or refine the types of questions that should be included in a particular questionnaire. The research instrument can then be prepared and a second pilot study conducted to evaluate other issues, such as clarity of instructions, clarity of questions, and so on. Alternatively, if the initial pilot study had started with an evaluation of the questionnaire, this may have revealed such an abundance of problems that it may be necessary to conduct a second pilot to ensure that all or most of the identified problems are properly addressed to minimize or avoid measurement error.

Pilot studies have been in use since the 1940s. They are suitable for both quantitative and qualitative studies and are helpful for determining suitability of instruments, data collection procedures, and sample population, to name a few. Additionally, pilot studies also serve other useful purposes. For example, pilot tests may help to convince funders that the particular study is worthy of funding. Or, in the case of a trial run, the pilot may reveal that the proposed relationship between variables may not exist, thus signaling to the researcher that the main study is no longer warranted. Pilot studies may even signal problems with local politics.

## Procedure

The sample population for a pilot study should closely mirror the intended targeted population. Thus, if the targeted population is university graduates, then the pilot participants should also be

university graduates. Generally, a convenience sample is used with about 50 to 100 participants—some studies may use even fewer participants. Compared to a pretest, a trial run or feasibility study is much more comprehensive because it is, in effect, a miniature version of the real research study. This means that it is executed using the same administration procedures that would be used to carry out the real study. Thus, if the research study intended to use interviews as the primary data collection method, the trial run should not use only experienced interviewers for the pilot because these interviewers would already be trained to deal with problematic situations.

If interviews are the primary data collection methodology, a pilot may focus on logistical issues such as interviewers' ability to follow survey directions, interviewers' ability to record data, time required to record responses, and so on. Thus, if the pilot study revealed that the average interview time exceeded the budgeted time, alternative strategies such as reducing the number of questions, using a smaller sample, or better training of interviewers may need to be considered to remedy this problem. The pros and cons of each strategy would need to be carefully considered because each strategy would have different implications for the research.

Likewise, if a self-administered questionnaire is being used, a pilot may focus on issues such as clarity of instructions for completing the survey, clarity of questions, and simplicity of questions, to name a few. Thus, if the pilot study revealed that Question 1 was ambiguous, Question 1 would need to be reworded. If most respondents circled—rather than shaded—their answers on the Likert scale, this would clearly indicate that the instructions were not clear. The instructions would need to be reworded; the researcher may decide to clarify the instructions by including an example of how responses should be shaded (e.g., **1 2 3 4**).

### Evaluation

Once the pilot study is completed, the next step is to evaluate the results. Interviewer debriefing is the most common form of evaluation. This process involves the researcher meeting with all of the interviewers in a group and discussing encountered problems. Often, standardized forms are used to record problems/deficiencies revealed from the pilot study. These forms may be completed by the interviewer, or, in the case of a pilot of a self-administered survey, the participant may be asked to fill out the form to provide general feedback on the instrument. Encountered problems are discussed in detail, and solutions are then identified to address any deficiencies. The research methodology is then refined.

### Problems With Pilot Studies

Pilot studies are undoubtedly helpful in identifying problematic issues in advance so that proactive action can be taken. Notwithstanding, a successful pilot study cannot guarantee success of the main study. During the actual study, other problems may be encountered that may not have been revealed by the pilot. For example, response rates may be much lower than anticipated, or contamination problems may surface if pilot participants are subsequently included in the main study. A more serious concern, however, may be if the research funding is terminated as a result of the pilot indicating that the study may no longer be original or warranted.

*Nadini Persaud*

***See also*** Follow-Up; Interviewing; Participants; Selection

#### Further Readings

Fink. A. (2003). *The survey handbook*. Thousand Oaks, CA: Sage.

Philliber, S. G., Schwab, M. R., & Sloss, G. S. (1980). *Social research: Guidelines to a decision-making process*. Itasca, IL: F. E. Peacock.

Van Teijlingen, E. R., & Hundley, V. (2001). The importance of pilot studies. *Social Research Update, 35*. Retrieved from http://sru.soc.surrey.ac.uk/SRU35.html

## PLACEBO

When conducting a double-blind clinical trial to evaluate a new treatment, the investigator is faced with the problem of what to give the control group. If there is no existing acceptable and

beneficial treatment against which the new treatment can be compared, the reasonable approach would be to give the control group no treatment at all, but this raises concerns that subjects may have hidden expectations or preferences that could influence the outcome. One solution is the use of a harmless "sham" treatment that is similar in all aspects such that the subjects and those administering the treatments cannot determine whether a subject is in the study group or the control group. This sham treatment is called the placebo. This entry discusses historical usage, difficulties with the placebo research design, and the placebo effect.

## Historical Usage

The term *Placebo Domino* . . . ("I shall please the Lord . . .") appears in a 5th-century translation of the Bible. By the end of the 18th century, *placebo* was used as a medical/pharmaceutical term, denoting a remedy designed to please the patient rather than effect any specific treatment. This sense of a deceptive yet morally acceptable therapy remained until the mid-20th century, when the true function of placebos came under critical scrutiny. Since 1955, the literature has generally advised caution in the use of the placebo, and has encouraged awareness of the limits of the effects of placebos, but their use has remained advisable when conducting randomized clinical trials.

## Difficulties With Placebo Design

Under ideal circumstances, both the active pharmaceutical treatment being evaluated and the placebo being used for the control group will be contained within similar gelatin capsules, tablets, or solutions, rendering them indistinguishable from each other as far as the participants are concerned. But when the active treatment under investigation has a distinct appearance or other characteristic such as strong taste, odor, or texture, or is a physical intervention, manipulation, or some invasive procedure, designing the placebo can present a challenge. How can the treatment and placebo be made indistinguishable from each other? And what if there are ethical issues for either group? Examples of each of these challenges are presented in this section.

### *Maintaining Similarity*

Most traditional Chinese medicine (TCM) remedies are often aromatic, dark-colored, bitter, and pungent. In a randomized crossover clinical trial of a TCM preparation for atopic eczema, all of these characteristics had to be simulated in the preparation of the placebo but without any therapeutic effect remaining. Participants who dropped out because of unpalatability were found to be just as likely in the placebo phase of the crossover trial as when the active treatment was being taken, suggesting that the organoleptic characteristics of the placebo were successfully indistinguishable from that of the active treatment.

### *Ethics*

In a study of the effectiveness of fetal brain cell implants for intractable Parkinson's disease, an ethical debate emerged over the use of nonimplant surgical control groups. Clearly, a theoretically perfect control group would have the presurgical preparation, opening of the skull, insertion of probes, and the closure procedure, but without the implant of fetal brain cells. Ethical objections were raised about the surgical risks for the control group, who would be without any potential benefit. But the need for the sham surgery was reinforced when it was learned that subjects receiving the sham procedure typically exhibited improvements in their Parkinson's symptoms for up to 6 months and were indistinguishable from patients who received the same surgery but with active (implant) treatment. The investigators were unable to attribute the observed improvements to either observer bias or the natural history of the disease.

## Placebo Effect

The placebo may produce a clinical effect by itself, arising perhaps from the participants' beliefs that the treatment should be effective, or from their subconscious preference to reveal a beneficial outcome from the treatment. This is known as the *placebo effect* and must be carefully considered when evaluating any effects that may be attributed to the treatment being investigated. The previous example is a curious case of a placebo effect in

which a substantial improvement was documented in the group that underwent surgery without implantation. It raises questions about exactly which component of the true surgical intervention constituted the effective treatment—was it really the implanted cells, or the neural stimulation of the surgical intervention, or some other component? Without the invasive and controversial placebo arm of the study, this question might never have been raised.

Determining the nature and scope of the placebo effect has long been subject to considerable controversy. Recent experimental results support two theories about placebo responses: *conditioning,* which proposes that changes in the body occur as a result of a stimulus that has been previously associated with active, positive processes; and *expectancy,* which proposes that changes occur in the body because the individual expects them to.

### *An Example of the Placebo Effect*

Irritable bowel syndrome is a common condition typically associated with consistent pain and bloating in the absence of any other obvious diagnosis. In a revealing study, a team led by Ted Kaptchuk did not provide any real treatment at all, but randomly divided 262 adults with IBS into three groups. The first group was told only that it was being studied and that outcomes were being collected. The second group received sham acupuncture (putting needles into the skin in a random way that is not based on the principles of acupuncture). The third group received sham acupuncture and an "enriched relationship" with the doctor. Moderate or substantial improvement in their IBS was reported by 3% of those patients in the observation-only group, 20% among the procedure-only group, and 37% in the group with procedure augmented by enriched communication from their physician. In the absence of any specific treatment, all of these results must be classified as placebo effects.

The evidence remains equivocal, however. In 2001, a meta-analysis of clinical trials with treatment, placebo, and no-treatment arms concluded that placebos have small or no effects. But a re-analysis in 2005 revealed that in many cases, the placebo effect was robust and delivers results close to the treatment effect.

*Timothy Sly*

***See also*** Double-Blind Procedure; Placebo Effect; Randomized Block Design

**Further Readings**

Freeman, T. B., Vawter, D. E., Leaverton, P. E., Godbold, J. H., Hauser, R. A., Goetz, C. G., & Olanow, C. W. (1999). Use of placebo surgery in controlled trials of cellular-based therapy for Parkinson's disease. *New England Journal of Medicine, 341,* 988–992.

Hróbjartsson, A., & Gotzsche, P. C. (2001). Is the placebo powerless? An analysis of clinical trials involving placebo with no treatment. *New England Journal of Medicine, 334,* 1594–1602.

Kaptchuk, T. J., Kelley, J. M., Conboy, L. A., Davis, R. B., Kerr, C. E., Jacobson, E. E., et al. (2008). Components of placebo effect: Randomized controlled trial in patients with irritable bowel syndrome. *British Medical Journal, 336,* 999–1003.

Macklin, R. (1999). The ethical problems with sham surgery in clinical research. *New England Journal of Medicine, 341,* 992–996.

Sheehan, M. P., & Atherton, D. J. (1992). A controlled trial of traditional Chinese medicinal plants in widespread non-exudative atopic eczema. *British Journal of Dermatology, 126,* 179–184.

Wampold, B. E., Minami, T., Tierney, S. C., Baskin, T. W., & Bhati, K. S. (2005). The placebo is powerful: Estimating placebo effects in medicine and psychotherapy from randomized clinical trials. *Journal of Clinical Psychology, 61*(7), 835–854.

## Placebo Effect

The placebo effect is any reaction to the administration of a placebo that is both clinically significant and salutary. A placebo is any treatment prescribed for a condition for which it is physically ineffective. The placebo effect has been successfully and repeatedly demonstrated on variables that are subjectively experienced, such as pain reduction, coping mechanisms, emotional well-being, and cognitions. The majority of research has centered on pain relief, also known as placebo analgesia. The magnitude of the effect and the percentage of those affected depend on a number of

factors, but research indicates that approximately one third of individuals who receive a placebo believing it to be a potent treatment report a significant analgesic effect.

The research involving objective measures of beneficence has been more controversial. Although some studies have described a modest effect of placebos on objective measures of healing, recent meta-analyses have indicated that these effects are, overall, nonsignificant. Objective measures of well-being and healing may be indirectly affected by positive changes in expectations for successful treatment or symptom relief. Although these are purely psychological variables, they may have very real consequences, because chronic pain, hopelessness, and a failure of usual coping strategies often occur during severe illness and are associated with a poor prognosis. In this entry, the history and elements of the placebo effect are described, and implications and ethical considerations for research are discussed.

## History of the Placebo Effect

The term *placebo* is a literal translation of the Latin phrase *I will please*. Its first documentation in an English dictionary did not occur until 1785. In 1920, T. C. Graves was the first to formalize the current conception of the medicine-like effects of placebo administration. In the 1950s, Henry Beecher attempted to quantify it. After a series of studies, he noted that the condition for which a patient was being treated greatly affected the percentage of those for whom a placebo alone was a satisfactory remedy. The effect, however, was much more prevalent than he expected, significantly influencing the amount of perceived pain in 30% of all patients.

A better understanding of the placebo effect spawned methodological transformations in the mid-20th century. Although placebo-controlled studies occurred on occasion throughout history, it was not until the 1930s that the widespread use of a dummy simulator was used as a means to test an experimental drug. This allowed researchers to compare the differences (or lack thereof) between the placebo effect of a specific medium on a specific condition and the drug effect on the same sample. Thus, experimenters could examine which drugs were significantly more effective than placebo treatment alone.

The inclusion of a placebo group is now standard practice in medical research. Drug manufacturers, prescribing physicians, and consumers demand assurance that medications or treatments are more effective than a placebo procedure. Therefore, before a treatment protocol or regimen is widely adopted, the placebo effect must be used as a baseline for such comparisons. Because the addition of a placebo group greatly complicates the research methodology and interpretation, the placebo effect is generally considered a necessary but burdensome aspect of medical research.

## Elements of the Placebo Effect

There are three components that affect the magnitude and prevalence of a placebo effect: patient characteristics, researcher manipulation, and the characteristics of the placebo treatment itself.

### *Patient Characteristics: Expectation and Desire*

The placebo effect is a psychosocial phenomenon, the result of social and individual expectations of the intended effect of a treatment. The expectation of benefit by the client is considered the most important and influential factor involved in the placebo effect. Knowledge that any treatment is a placebo will, generally, nullify any placebo effect. This seems an obvious but salient point: Without hopeful expectations, a client will not garner the benefits of the placebo. In fact, it appears that when patients are administered real analgesics without their knowledge, they benefit significantly less than those who are openly administered the drug.

The central building block of patient expectancy occurs through classical conditioning. Medicinal placebos may be especially effective, because most people have had a history of being medicated effectively. An unconditioned stimulus (an active intervention) has been paired repeatedly with an unconditioned response (benefit). The active intervention is then mimicked by the placebo intervention, resulting in the conditioned response of benefit. Conditioning has produced lasting placebo effects under falsified laboratory settings and through naturalistic field observation.

Specific levels of pretreatment expectations of posttreatment pain account for between 25% and 49% of the variance associated with that experienced after a placebo analgesic administration. Furthermore, patients have a tendency to overestimate the amount of pretreatment pain and underestimate the amount of posttreatment symptom reduction. The degree to which these memory distortions occur is highly correlated to pretreatment expectations of success.

Desire for an effective treatment can also create or enhance placebo effects. Although every patient seeks treatment because he or she desires to be free of symptoms, such avoidance goals are often not powerful enough to marshal a noticeable placebo effect.

Increasing the desirability of a placebo response can greatly enhance the effect. For instance, participants who believe particular responses to a medication are indicative of positive personality characteristics are more likely to report those responses than those who were not similarly motivated. Additional studies have determined that very subtle, nonconscious goals can influence the magnitude of the placebo effect. The results indicated that one will likely be influenced by a placebo if it suits his or her goals, even if those goals are not overtly conscious.

Finally, patients in great pain have an increased desire for any treatment to be effective. The greater the pain, the more likely a placebo response will occur.

### *Researcher Manipulation: Suggestion and Credibility*

The researcher can drastically influence patient expectations of treatment through making suggestions and enhancing credibility. Whereas suggestions usually occur on an overt, explicit basis, credibility is often a given of the therapeutic experience, both implicit and understood.

In Western society, there are certain symbols of healing credibility. Having one's own practice, wearing a white lab coat, displaying diplomas on the wall, and having a PhD behind one's name are all ways to establish expectancy through the appearance of expertise. When these culturally relevant tokens are assumed, the researcher becomes sanctioned as a healer, and the resultant credibility mobilizes the patient's expectancy and results in greater therapeutic gain.

The researcher can also enhance his or her trustworthiness and attractiveness to maximize credibility. An attractive, socially sanctioned healer is likely to inspire motivation and hope within the patient, and is therefore more likely to elicit a placebo response.

A number of studies have examined the effect of overt verbal suggestions made by physicians and the client's resulting placebo effect. Direct allusions to the potency of a treatment are effective means of creating expectancies in patients and are easily manipulated by the researcher. Research indicates that the emphasis, tone, and content of even a small comment can greatly affect the magnitude of a client's placebo response.

### *Placebo Characteristics: Cultural Association With Potency*

The characteristics of the placebo treatment itself can influence its effects. More accurately, the cultural significance of the method of placebo delivery can enhance its effectiveness. Apparent invasiveness and strength of the intervention are ordinally related to the strength of the placebo effect for that mode of intervention. For instance, most Americans expect a larger benefit from surgery than from only a single injection from a hypodermic needle. Most Americans also expect more benefit from that same injection than from a simple pill. An analysis of the amount of placebo-induced benefit from each form shows the same trend: The placebo effects for one sham surgery are greater than those for a single injection, which are greater than those from a single pill, even if they all promise the same results.

## Implications for Research

The placebo effect occurs during every medical or psychological intervention into which a participant carries an expectation or motivation for success. To some extent, it affects every participant. It is not possible to determine how the elements of the placebo effect will interact, and it is therefore impossible to predict to what extent an individual will exhibit a placebo response. Given this unpredictability, the placebo effect must be taken into

account and quantified to accurately assess the efficacy rates of a treatment.

To test the efficacy of a particular drug on a particular condition, participants are often randomly assigned into one of three conditions: the experimental group (E), the placebo group (P), and the natural history group (NH). The experimental group receives the active drug. The placebo group receives an inert substance through a procedure identical to that used to administer the experimental drug. In double-blind procedures, neither the researchers nor the participants know which participants belong to the placebo-control group and which belong to the experimental-active group. This practice eliminates any conscious or unconscious experimenter bias that creates expectations that may differentially inflate the placebo effect. Finally, the natural history group does not receive any intervention at all. Participants in this group are often wait-listed and monitored until the experimental trial is complete.

By creating fomulas that assess the differences between the three groups, researchers can quantify the magnitude of the placebo effect, the efficacy of the active drug, and the overall effect of the treatment. The outcome of the natural history group provides an absolute baseline (NH). This represents, on average, the course of the condition when untreated. Individuals in this group are susceptible to using methods that differ from the standard treatment of the other two groups, such as homeopathy, vitamins, or exercise. Such extraneous variables should be considered and carefully controlled.

The strength of the placebo effect (PE) is determined by examining the differences between the natural history group and the placebo group (PE = P – NH). This fomula indicates the extent to which the elements of a placebo have influenced the subjective responses of some participants.

The efficacy of the experimental treatment can be determined two different ways. The absolute efficacy (AE) is determined by the difference between the experimental group and the natural history group (AE = E – NH). This fomula represents the total effect of the treatment. Although it is important that any treatment be superior to no treatment at all, this is usually not the figure of interest. Any inert substance, if properly administered, can achieve some level of success (the placebo effect). The active component (AC) efficacy is the difference between the experimental group and the placebo group (AC = E – P). This fomula accounts for the placebo effect and is therefore the standard by which the efficacy of any treatment is judged.

Because placebo-controlled studies involve the administration of a physically inactive substance, treatment modes that do not require physical interaction, such as psychotherapy, have not previously been viewed as amenable to placebo research. There is a great deal of controversy surrounding what can be considered a placebo in noninvasive treatments, and how that definition can be used in outcome research.

## Ethical Considerations of Placebo Research

There are a number of ethical considerations when using placebos in research. First, the use of placebos is considered deception. Research ethicists and legal experts generally advocate a conservative use of placebos in research to protect researchers and clinicians from unwarranted ethical violations and litigation. They condone their use only under certain conditions: There is no efficacious alternative therapy, the research will result in practical benefit, the duration of the trial will be short, there is little risk that the participant's condition will worsen, the participant has been fully apprised as to the state of the biomedical knowledge, and the participant has signed a fully informed consent that complies with the applicable legal and ethical codes. Furthermore, once the study has begun, it may become unethical to continue the clinical trial once an experimental treatment has shown definitive evidence of efficacy. If this occurs, researchers are obligated to inform participants in the placebo group (and any other comparison groups) and offer them the experimental treatment.

In general, placebo-controlled randomized trials are complicated, require vast resources, and are ethically and legally perilous. They can provide valuable information but are appropriate only under specific conditions and should not be attempted haphazardly.

*Thomas C. Motl*

*See also* Clinical Trial; Control Group; Double-Blind Procedure; Informed Consent; Placebo

### Further Readings

Beecher, H. K. (1955). The powerful placebo. *Journal of the American Medical Association, 159,* 1602–1606.

Borkovec, T. D., & Sibrava, N. J. (2005). Problems with the use of placebo conditions in psychotherapy research, suggested alternatives, and some strategies for the pursuit of the placebo phenomenon. *Journal of Clinical Psychology, 61*(7), 805–818.

De Deyn, P. P., & Hooge, R. (1996). Placebos in clinical practice and research. *Journal of Medical Ethics, 22*(3), 140–146.

Kapp, M. B. (1983). Placebo therapy and the law: Prescribe with care. *American Journal of Law & Medicine, 8*(4), 371–405.

Moerman, D. E. (2002). The meaning response and the ethics of avoiding placebos. *Evaluation & the Health Professions, 25,* 399–409.

Price, D. D., Finniss, D. G., & Benedetti, F. (2008). A comprehensive review of the placebo effect: Recent advances and current thought. *Annual Review of Psychology, 59,* 565–590.

Shapiro, A. K., & Shapiro, E. (1997). *The powerful placebo: From ancient priest to modern physician.* Baltimore: Johns Hopkins University Press.

## Planning Research

Planning research is an applied investigation that uses empirical observations in the development and assessment of plans or planning inquiry. It is an "applied research" strategy in that the planner investigates everyday situations that are commonly faced by local residents. The main purpose for research in planning is to expand on the breadth, depth, and applicability of what the planner knows about a particular topic to either serve as a foundation or improve the suitability and success of his or her developed plans.

Accessing multiple data sets and organizing research projects around a mixed-method investigative strategy is generally standard practice in planning research. This is due to the multifaceted nature of the problems addressed that requires the planner to access different qualitative and quantitative data sets in order to generate a holistic understanding about the problem at hand. Qualitative data sets that are commonly generated in planning research projects include spoken comments from local residents, aerial and eye-level photographic images, and field observations. Quantitative data sets commonly used by a planner include census and demographic statistics, built and nonbuilt land use information, economic data, and vehicle traffic counts.

In addition, taking on complex problems is also part of the "systems approach" of how a planner looks at cities and the environment that directly influences how he or she frames research questions. Here, a region, city, and neighborhood are understood as being made up of a series of interconnected systems (social, economic, political, transportation, environment, etc.). A change in one system will have a direct impact on another system. For example, a dramatic increase in the population will have noticeable impacts in housing, health care, education, economy, and transportation. Furthermore, using the systems approach to frame planning problems provides the opportunity for the planner to identify potential solutions in one system through adjustments made in another system.

This entry examines two major junctures of the planning process: research and evaluation. The role of the community and politics is also examined.

### Research in Planning

Planning is normatively structured as a linear decision-making process that makes plans as a way to achieve desired outcomes. The planning process is anchored by five sequential steps: (1) problem/opportunity, (2) research, (3) policy/programming, (4) implementation, and (5) evaluation. The plan-making process begins when a planner is presented with a problem or opportunity. Here, the planner establishes the anatomy of the problem/opportunity and determines if the development of a plan is the best way to achieve a desired goal. The research phase in the planning process begins with a need for more information ("research question") to shore up any deficiencies in his or her knowledge about the planning task being addressed. Research most commonly occurs during the research and eval-uation junctures (Steps 2 and 5) in the planning process. However, the purposes for

research in these two planning junctures are very distinct. Investigations made during the research phase of the planning process can take one of two directions: research as part of a comprehensive planning process or strategic research to investigate a specific planning problem.

### *Comprehensive Planning Research*

Comprehensive planning is a long-term plan-making process that strives to develop the best possible solution based on an exhaustive review (within reason) of multiple data sources. It takes an integrative, multivariable research approach with the goal of generating a comprehensive understanding of both the problem at hand and its embedded context. A city developing a comprehensive plan to determine future planning directions is an example of comprehensive planning.

Planning research within the comprehensive planning process looks at a fairly standard list of secondary data sources to provide a baseline inventory assessment of existing community conditions. According to Eric Damian Kelly and Baroverlinea Becker, this list of comprehensive planning research topics includes population, housing, economy, land use, natural environment, resources and sustainability, circulation, community services, and parks and recreation. Given the broad scope of topics covered in the comprehensive planning research investigation, the planner gathers mostly secondary data sources to expedite the compre-hensive planning process. Whenever possible, the planner analyzes the data longitudinally to determine if any significant trends are developing among individual data sets (e.g., increase in automobile traffic) and comparatively to see if any cumulative relationships are developing between data sets (e.g., an increase in total population and in elementary school classroom enrollment). He or she also looks to the future with projections (e.g., population or traffic projections) to estimate possible developments. Conclusions drawn from these baseline inventory assessments help the planner to identify problems and opportunities and their magnitude that will provide empirical direction in the community's development of goals and objectives in the following steps in the planning process.

Most of the data sets used in the comprehensive planning process are spatially defined, which allows the planner to store, analyze, and display through a series of maps via a GIS (geographic information system) analysis. Commonly known as a land use inventory, these data sets are used to make a series of maps that document a host of land use characteristics that include zoned land use (commercial, industrial, residential); public land; open space; topography; flood plains; census data; water and sewer infrastructure; school districts; traffic corridors; and public transportation routes. Through the use of GIS mapping, physical land use data (mountains and valleys) are visually presented through satellite imagery.

### *Strategic Planning Research*

Strategic planning research is employed when the planner sets out to investigate a specific problem. It is more of a short-term plan-making process in comparison to comprehensive planning. Strategic planning is action-oriented and looks more to solving problems and making targeted advancements than developing the best possible solution. A city addressing the problem of traffic congestion at an intersection is an example of strategic planning.

Even though strategic planning is very focused, the planner continually addresses the problem as part of a larger system of intervening issues. For example, recent population increases have resulted in local residents requesting that the city build more soccer fields to meet the growing interest in youth soccer. To determine if the provision of new soccer fields is possible, strategic planning research is needed to help the city make an informed decision. In this research project, the planner needs to develop a mixed-method research design that identifies all the necessary data sets to allow them to generate a holistic understanding around the need for new soccer fields. In this example, the planner can design the research project around a sequential series of three interconnected data sets. The first data set will assess the supply of existing soccer fields in relation to a census analysis to analyze the longitudinal growth in the population and estimate the future demand for more soccer fields. This could be followed by an assessment of the supply of available suitable parcels of land and their

proximity to where people live and anticipated residential growth. Finally, an impact analysis of the proposed soccer sites and how they will affect neighboring residents, traffic systems, and the environment will need to be completed.

## Community-Based Data

What is unique in planning research, in comparison to other types of applied research investigations, is that the community plays a very important role in terms of being a source of data as well as providing an assessment of validity. Planners actively look to the community as an important source of data. This is ontologically based on John Dewey's "immediate empiricism," where everyone in the community is able to experience events and empirically report what they observed.

Planners systematically draw on community observations either through surveys, small-group qualitative investigations such as focus group research projects, or independent citizen reports. Mailed surveys are one way for planners to gain input from large numbers of people in the community. A planner can use a mail survey during a comprehensive planning project to assess future interest in the development of new parks. Focus group studies are good for problems that have distinct subcommunities that experience a shared situation but from different perspectives, and the planner is interested in getting more detailed qualitative observations in the community's own words. In the strategic planning research project to add more soccer fields, how neighboring residents experience the new fields will be very different from how the people who will use the soccer fields experience them, and the planner can use a focus group research strategy to discover this. Individual community members volunteering field observations has been growing in popularity in the past decade, especially among nature and wildlife organizations. The precedent for this type of organized community observation comes from the U.S. National Weather Service, dating back to the 1800s with citizens regularly observing and recording weather activity. Organizations like the U.S. National Park Service use citizen field observations in their database research.

## Community Face Validity

Planning is very much a community decision-making process. At the completion of all the explorations during the research phase, and once planning recommendations have been made at the end of the policy/programming phase, the proposed plan goes to the community for an assessment and usually a vote. The community plays an important role in determining planning priorities as well as evaluating the effectiveness of implemented plans. Throughout the entire planning process, community members are consulted regularly for their input. They assess the actions and progress of the planning process either formally, through city council meetings and public hearings, or informally, with neighborhood informational meetings and open houses. In planning research, the community provides a type of "face validity" assessment of planning research conclusions. Individuals in the community make commonsense assessments of research conclusions and ask, "On the face of things, does this research make sense?" Some planning research projects, such as those carried out during an environmental impact statement, schedule several public involvement meetings throughout the planning research process.

## Program Evaluation

After proposed plans are agreed upon by the community and have been implemented, the second major research juncture begins during the evaluation phase in the planning process. Program evaluation is a comparative research process that uses empirical observations of existing baseline observations of an existing state of affairs to assess the impact (success or failure) of an implemented program. Program evaluation is divided into two types of investigations: formative evaluation that assesses the process of how programs are implemented, and outcome/summative evaluation that examines the post facto results produced at the completion of the program. Planning program evaluation research mostly focuses on outcome evaluations.

Two of the more common types of planning outcome evaluation research projects are cost-benefit and cost-effectiveness evaluations. Cost-benefit analysis is the comparison of a program's costs to its measurable benefits. Both costs and

benefits are measured in the same unit, usually money. Cost-benefit assessments are commonly used to assess the efficiency of a program. Lower costs than identified benefits means the program is efficient. Public transportation planning commonly uses cost-benefit analysis in comparing the costs of providing public transit services to the revenue generated through its patrons as one way to assess the success of public transit services. Cost-effectiveness analysis uses the same comparison process as in a cost-benefit analysis, but "costs" and "unit of service" as measures of effectiveness are recognized in two different units. An example of a cost-effectiveness planning research project is assessing the effectiveness of a park in terms of number of visitors in comparison to the costs to provide and maintain the park.

### Politics and Planning Research

Planning research usually takes place within a politically charged environment. The combination of applied research orientation in the investigation of everyday problems, active insertion of community comments throughout the plan-making process, and the struggle over scarce resources among competing interest groups can and does politicize planning research. Research conclusions derived from very defendable investigations can have extremely negative ramifications, especially with politically disenfranchised populations and the environment. An example of this is the U.S. federally funded urban renewal projects during the 1950s and 1960s that adhered to a detailed research strategy that produced planning recommendations to wipe out entire low-income, multi-ethnic communities to make way for new upscale commercial developments. Some of the more pressing challenges for planning researchers are to ensure that they execute defendable research in terms of the criteria of validity and reliability as well as ensure that the research outcomes are equitably and ethically fair to everyone in the community and the environment.

*John Gaber and Sharon L. Gaber*

***See also*** Focus Group; Mixed Methods Design; Qualitative Research; Quantitative Research; Secondary Data Source

**Further Readings**

Friedmann, J. (1987). *Planning in the public domain: From knowledge to action*. Princeton, NJ: Princeton University Press.

Gaber, J., & Gaber, S. (2007). *Qualitative analysis for planning and policy: Beyond the numbers*. Chicago: APA, Planners Press.

Kelly, E. D., & Becker, B. (2000). *Community planning: An introduction to the comprehensive plan*. Washington, DC: Island Press.

Laws, S., Harper, C., & Marcus, R. (2003). *Research for development: A practical guide*. London: Sage.

Levy, J. (2009). *Contemporary urban planning* (8th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

## Poisson Distribution

The Poisson distribution is a discrete probability distribution that is often used for a model distribution of count data, such as the number of traffic accidents and the number of phone calls received within a given time period. This entry begins with a definition and description of the properties of the Poisson distribution, which is followed by a discussion of how the Poisson distribution is obtained or estimated. Finally, this entry presents a discussion of applications for the distribution and its history.

### Definition and Properties

The Poisson distribution is specified by a single parameter $\mu$ which determines the average number of occurrences of an event and takes any positive real number (i.e., $\mu > 0$). When a random variable $X$, which can take all nonnegative integers (i.e., $X = 0, 1, 2, \ldots$), follows the Poisson distribution with parameter $\mu$, it is denoted by $X \sim \text{Po}(\mu)$. The probability mass function of $\text{Po}(\mu)$ is given by

$$P(X = x) = \frac{\mu^x e^{-\mu}}{x!},\ x = 0, 1, 2, \ldots,$$

where $e$ is the base of the natural logarithm ($e = 2.71828\ldots$). The mean of $X$ is $\mu$, and the variance is also $\mu$. The parameter $\mu$ is sometimes decomposed as $\mu = \lambda t$, where $t$ is the length of a given time interval and $\lambda$ denotes the "rate" of occurrences per unit time.

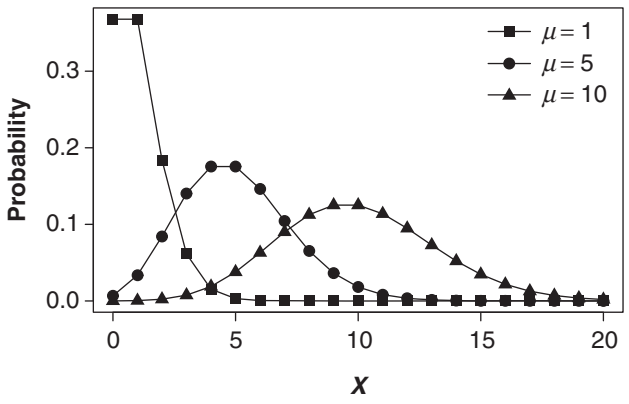



**Figure 1** Probability Mass Functions of the Poisson Distributions With $\mu = 1$, 5, and 10

Figure 1 depicts the probability mass functions of the Poisson distributions with $\mu = 1, 5$, and 10, respectively. The horizontal axis represents values of $X$, and the vertical axis represents the corresponding probabilities. When $\mu$ is small, the distribution is skewed to the right. As $\mu$ increases, however, the shape of distribution becomes more symmetric and its tails become wider.

The Poisson distribution arises as a result of the stochastic process called the *Poisson process,* which builds upon a set of postulates called the *Poisson postulates*. Roughly speaking, the Poisson postulates posit the following conditions on the probability of events occurring in given time intervals: (a) The numbers of events occurring in nonoverlapping time intervals are independent, (b) the probability of exactly one event in a very short (i.e., infinitesimally short) time interval is approximately proportional to the length of the interval, (c) the probability of more than one event in a very short interval is much smaller than the probability of exactly one event, and (d) the above probability structure is unchanged for any time interval. These postulates lead to a set of differential equations, and solving these equations for the probability of the number of events occurring in a given time interval produces the Poisson distribution.

The way by which the Poisson distribution is derived provides a basis for the distribution to be used for a model of rare events; the Poisson postulates state that the event occurs rarely within each small time interval but is given so many opportunities to occur. In this sense, the Poisson distribution is sometimes called *the law of small numbers*.

## Relationship to Other Probability Distributions

The Poisson distribution is obtained as a limiting distribution of the *binomial distribution* with probability $p$ and the number of trials $n$. Let $\mu = np$, and increase $n$ to infinity while $\mu$ is held constant (i.e., $p$ approaches 0). The resulting distribution is Po($\mu$). This fact, in turn, implies that the Poisson distribution can be used to approximate the binomial distribution when $n$ is large and $p$ is small. This is of great use for calculating binomial probabilities when $n$ is large, because the factorials involved in the binomial probability fomula become prohibitively large.

If the number of occurrences of an event in a given period of time $t$ follows Po($\lambda t$), then time $T$ between two successive occurrences (which is called the waiting time) is also a random variable and follows the *exponential distribution* with mean $1/\lambda$. Thus, if occurrences of an event in a given time interval $t$ follow the Poisson distribution with rate $\lambda$, then the waiting time follows the exponential distribution and the average waiting time for the next event to occur after observing an event is $1/\lambda$

When $\mu$ is large, Po($\mu$) is well approximated by the normal distribution N($\mu$,$\mu$) because of the central limit theorem. Thus, if $X \sim$ Po($\mu$), then $Z = (X - \mu)/\sqrt{\mu}$ approximately follows the standard normal distribution N(0,1), from which probabilities regarding $X$ are obtained easily. For the probability calculation, the chi-square distribution can also be used; the probability that $X \sim$ Po($\mu$) takes a value less than or equal to $x$ is given by $P(X \leq x) = P(\chi^2_{2(x+1)} > 2\mu)$, where the right-hand side denotes the probability that the chi-square random variable with $2(x + 1)$ degrees of freedom takes a value greater than $2\mu$.

## Estimation

The parameter $\mu$ can be estimated by the sample mean of a random sample $X_1, X_2, \ldots, X_n$ from Po($\mu$). That is, $\hat{\mu} = \overline{X} = \sum_{i=1}^{n} X_i/n$. It is an unbiased estimator with the minimum variance,

which is $\mu n$, as well as the maximum likelihood estimator.

For the interval estimation, approximate $100(1-\alpha)\%$ confidence limits $(\mu_L, \mu_U)$ for $\mu$ can be found by

$$\mu_L = 0.5\chi^2_{2x,\alpha/2}, \mu_U = 0.5\chi^2_{2(x+1),1-\alpha/2},$$

where $x$ is an observed value of $X \sim \text{Po}(\mu)$, and $\chi^2_{df,p}$ is the value of the chi-square variable with degrees of freedom *df,* which gives the lower-tail probability $p$. If $\mu$ is expected to be large, then the normal approximation may be used as well.

## Applications

### *Poisson Regression/Loglinear Models*

Poisson regression is a regression model used for count data. Especially when used for contingency tables, it is called the *loglinear model*. Poisson regression models are fomulated as *generalized linear models* in which the canonical link function is the log link and the Poisson distribution is assumed for the dependent variable.

In the standard case in which there are $K$ linear predictors $x_1, x_2, \ldots, x_K$ and the dependent variable $Y$, which represents a count given $x$s, the Poisson regression model is expressed by the equation

$$\ln[\mu(x)] = b_0 + b_1x_1 + \cdots + b_Kx_K,$$

where ln is the natural logarithm, $\mu(x)$ is the expected count of $Y$ given $x_1, x_2, \ldots, x_K$, and $b_0$, $b_1, \ldots, b_K$ are the regression coefficients. The distribution of count $Y$ given $x_1, x_2, \ldots, x_K$ is assumed to be $\text{Po}[\mu(x)]$, where the log of the expected value $\mu(x)$ is determined by a linear combination of predictors $x_1, x_2, \ldots, x_K$. The regression coefficients are estimated from $N$ sets of observed data $(Y_i, x_1, x_2, \ldots, x_K)$, $i = 1, 2, \ldots, N$.

When applied to a contingency table, the above linear Poisson regression model is equivalent to the loglinear model, that is, a model for expected cell frequencies. Suppose that the table comprises two factors, $A$ (row factor) and $B$ (column factor), whose levels are denoted by $i = 1, \ldots, r$ and $j = 1, \ldots, c$, respectively. Let $Y_{ij}$ be the frequency of the $(i, j)$ cell of the table and $\mu_{ij}$ be the corresponding expected cell frequency. The Poisson loglinear model assumes that $Y_{ij} \sim \text{Po}(\mu_{ij})$, and

$$\ln(\mu_{ij}) = \lambda + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB},$$

where $\lambda$ is the grand mean, $\lambda_i^A$ and $\lambda_j^B$ are the main effects of the $i$th row and the $j$th column, respectively, and $\lambda_{ij}^{AB}$ is the interaction for the $(i, j)$ cell, as in the usual two-way ANOVA model. The above model is called the saturated model, by which the observed cell frequencies are completely reproduced by the estimates of $\mu_{ij}$. Then, various hypotheses can be tested by setting certain effects to zero. For example, setting all interactions to zero (i.e., $\lambda_{ij}^{AB} = 0$ for all $i$ and $j$) leads to the model of independence, that is, there is no association between Factors $A$ and $B$, and each cell frequency can be estimated from its marginal frequencies. Further setting all the $A$ main effects to zero (i.e., $\lambda_i^A = 0$ for all $i$) implies that the cell frequencies do not differ by different levels of $A$. A model is tested by comparing its goodness of fit with a more general model. For example, the independence model can be tested against the saturated model. If the goodness of fit of the independence model is not significantly worse than that of the saturated model, the independence model is accepted.

### *Poisson Process Model for the Number of Misreadings*

Frederick Lord and Melvin Novick introduced several measurement models, which were originally proposed by Georg Rasch, that were based on Poisson distributions. One of the models was intended for the number of misreadings of words in an oral reading test, which Rasch considered measures one aspect of reading ability. As in other measurement models that Rasch proposed, such as the one-parameter logistic item response model (i.e., the Rasch model), his intention was to estimate individual ability separate from the difficulty of texts. This is the key feature of his Poisson process model as well.

Suppose that there are several texts, each of which is denoted by $g$ and consists of $N_g$ words. Rasch assumed that the probability that person $i$ misreads a given word in text $g$ is a ratio of the text's difficulty, $\delta_g$, to the person's ability, $\xi_i$, that is, $\theta_{gi} = \delta_g/\xi_i$, where $\theta_{gi}$ is the probability of misreading of a given word in text $g$ by person $i$. The

text difficulty parameter $\delta_g$ takes a value between 0 and 1, with the most difficult text having $\delta_g = 1$. The person ability parameter takes a value from 1 to infinity, with the least able person having $\xi_i = 1$ (this is a hypothetical person who certainly misreads any word). These constraints are required to uniquely determine values of the parameters. Then, with a sufficiently long text, a small probability of misreading, and independence of individual misreadings given the person's ability, the number of misreadings in text $g$ by person $i$ follows the Poisson distribution with parameter $\lambda_{gi} = N_g\theta_{gi} = N_g\delta_g/\xi_i$, which represents the expected number of misreadings in text $g$ by person $i$.

After observing a data matrix, which consists of $x_{ig}$, the number of misreadings by person $i$ ($= 1, \ldots, N$) and text $g$ ($= 1, \ldots, G$), we can estimate relative sizes of the text difficulty parameters $\delta_g$ and person ability parameters $\xi_i$ separately as if they were the main effects in two-way ANOVA. Then, the estimates are rescaled so that they satisfy their constraints.

### History

Siméon-Denis Poisson, a French mathematician and physicist, discovered the Poisson distribution in the 19th century (although Abraham de Moivre provided the same result even earlier in the 18th century). In his work in probability theory, Poisson is also known for his introduction of the term *the law of large numbers* (*la loi des grands nombres* in French), which was originally called *Bernoulli's theorem* after its original discoverer Jakob Bernoulli. Poisson found the distribution by considering the limit of the binomial distribution as described above. He applied it to modeling the deliberations of juries, but it did not receive popularity at that time. Later, the Poisson distribution drew the attention of several mathematicians, such as Ladislaus Bortkiewicz, who gave the name the law of small numbers to the distribution; Thorvald Thiele; and William Gosset. Now the Poisson distribution is even considered as a standard distribution for a discrete random variable, as the normal distribution is the standard distribution for a continuous random variable.

*Kentaro Kato and William M. Bart*

***See also*** Bernoulli Distribution; Loglinear Models; Normal Distribution; *Probabilistic Models for Some Intelligence and Attainment Tests;* Probability, Laws of

### Further Readings

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York: Wiley.

Johnson, N. L., Kemp, A. W., & Kotz, S. (2005). *Univariate discrete distributions* (3rd ed.). Hoboken, NJ: Wiley.

Lindgren, B. W. (1993). *Statistical theory* (4th ed.). Boca Raton, FL: Chapman & Hall/CRC.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

## Polychoric Correlation Coefficient

The polychoric correlation coefficient, $\rho$, is used for correlation when the data consist of observations from two ordinal variables, each having an arbitrary number of response categories. Strictly speaking, the polychoric correlation coefficient estimates the correlation between two unobserved bivariate normal variables assumed to underlie the observed ordinal variables. The polychoric correlation coefficient is a generalization of the tetrachoric correlation coefficient, a statistic used to estimate correlation based on two dichotomous variables. Under assumptions, the polychoric correlation provides a correlation estimate that is entirely free of the attenuation caused when two normally distributed variables are "crudely categorized"—that is, when they are reduced to sets of ordinal categories. This attenuation in the correlation estimate leads to bias in parameter estimates when the biased correlations are used in methods such as factor analysis or structural equation modeling. By contrast, parameter estimates from analysis of polychoric correlations tend to be unbiased. Although the polychoric correlation coefficient is built upon an assumption of underlying bivariate normality, simulation studies show that the polychoric correlation coefficient is somewhat robust to violations of this assumption. The polychoric correlation coefficient will tend to yield a stronger (in absolute value) correlation estimate than a Pearson product-moment

correlation applied to ordinal variables, especially when the number of response categories for each variable is small (less than five) and when the distributions of the ordinal variables are skewed. Yet the polychoric correlation coefficient, which traces its origins to Karl Pearson's work in the early 1900s, shares the same expected value as the Pearson product-moment correlation. By contrast, nonparametric correlation coefficients such as Spearman's rank-order correlation and Kendall's $\tau_b$ have different expected values, making them less attractive as substitutes for the Pearson product-moment correlation.The polychoric correlation coefficient has been used prominently for factor analysis and structural equation modeling of ordinal data. The statistic has definite advantages over some alternative approaches but also has substantial drawbacks. However, continuing innovation in structural equation modeling, as well as problems with the use of the polychoric correlation for this purpose, seem likely to make this application of the polychoric correlation coefficient less prominent in the future.

## Estimating the Polychoric Correlation Coefficient

Imagine two variables $X$ and $Y$, which might represent responses to two items on a questionnaire, where those responses were limited to a set of ordered and mutually exclusive categories. The items might be two Likert scale items, each with a set of response categories labeled *strongly disagree, disagree, neither agree nor disagree, agree,* and *strongly agree*. Researchers will often assign numbers to these categories, such as 1, 2, 3, 4, and 5. A researcher who then uses these numbers to compute statistics—such as a Pearson product-moment correlation between $X$ and $Y$—is implicitly assuming that the variables $X$ and $Y$ have at least interval scale. If the variable has only ordinal scale, however, then the specific numbers assigned to different categories signify only the ordering of the response categories—they cannot be used for computation. Unlike the Pearson product-moment correlation, the polychoric correlation is derived, not computed from the response category scores.

The logic of the polychoric correlation coefficient assumes that the observed ordinal variable $X$ is a categorization of an underlying continuous, normally distributed variable, $X^*$. The $g$ response categories of $X$ result when the normal distribution of $X^*$; is divided by $g - 1$ unknown thresholds $\tau_i$:

$$\begin{aligned} X &= \alpha_1 \text{ if } -\infty \le X^* < \tau_1 \\ X &= \alpha_2 \text{ if } \tau_1 < X^* < \tau_2 \\ &\cdots \\ X &= \alpha_g \text{ if } \tau_{g-1} < X^* \le \infty. \end{aligned}$$

Similarly, the $h$ categories of $Y$ result from categorization of a normally distributed $Y^*$:

$$\begin{aligned} Y &= \beta_1 \text{ if } -\infty \le Y^* < \tau_1 \\ Y &= \beta_2 \text{ if } \tau_1 < Y^* < \tau_2 \\ &\cdots \\ Y &= \beta_h \text{ if } \tau_{h-1} < Y^* \le \infty. \end{aligned}$$

Estimating the polychoric correlation, then, involves estimation of $(g - 1) + (h - 1) + 1 = g + h - 1$ parameters (the thresholds plus the correlation) from the $g \times h$ cross-tabulation of $X$ and $Y$. The log likelihood for a simple random sample of size $n$ from an infinite population is

$$\ln L = \ln C + \sum_i^g \sum_j^h n_{ij} \ln \pi_{ij},$$

where $C$ is a constant, $n_{ij}$ is the number of observations in cell $i,j$ and $\pi_{ij}$ is the probability of a single observation falling into cell $i,j$. Given the underlying assumption of bivariate normality, then

$$\begin{aligned} \pi_{ij} = {} & \Phi_2(i,j) - \Phi_2(i-1,j) - \Phi_2(i,j-1) \\ & - \Phi_2(i-1,j-1), \end{aligned}$$

where $\Phi_2$ is the bivariate normal distribution function with correlation $\rho$.

### *Alternative Approaches to Estimation*

In practice, researchers have developed at least three major approaches to estimation. In the early 20th century, in the pre-computer era, scholars commonly estimated the thresholds from the marginal counts of the cross-tabulation table. With the scale of the ordinal variables being arbitrary,

threshold values were chosen from the standard normal $Z$ distribution, based on the proportion of responses in each category. For example, if 2.5% of responses to $X$ were in the lowest response category, the researcher would set a threshold value of –1.96 between the lowest category and the next lowest. If 47.5% of responses fell into the next response category, then the next threshold value would be 0.00. With the threshold values fixed, the researchers would then estimate the polychoric correlation by setting the derivative of the likelihood function with respect to the correlation equal to zero:

$$\frac{d\ln L}{d\rho} = 0,$$

then solving for the correlation. An alternative approach, made more feasible with computer assistance, is to estimate all $g + h - 1$ parameters simultaneously via maximum likelihood. Differences between these two approaches tend to be small, which recommends the older, faster, two-stage method. Either way, estimates of the polychoric correlation tend to be unbiased when the observed distribution is symmetric or with moderate levels of skew.

Both of these methods estimate each polychoric correlation one at a time. Researchers with a number of observed variables may actually want a matrix of polychoric correlations. When the correlations are estimated in this pairwise fashion, however, there may be inconsistencies across the matrix, leading to a correlation matrix that is not well-conditioned for multivariate analysis. This is particularly a problem for researchers who want to apply factor analysis or structural equation modeling (SEM) to the matrix of polychoric correlations. However, simultaneous estimation of an entire matrix of polychoric correlations, assuming an underlying multivariate normal distribution, leads to complex integrals. Researchers have attempted to make this problem more tractable through reparameterization, but this multivariate approach may still be computationally infeasible for moderately sized models, even with modern computing power. As a result, most current software uses a bivariate approach when estimating polychoric correlations.

## Polychoric Correlations in Structural Equation Modeling

However, using polychoric correlations in structural equation modeling of ordinal variables raises further problems. Maximum likelihood estimation of a structural equation model assumes that the elements of the covariance matrix being analyzed follow a joint Wishart distribution. This assumption involves only limited negative consequences when researchers analyze matrices of Pearson product-moment correlations. However, the elements of a matrix of polychoric correlations do not follow a joint Wishart distribution. Although parameter estimates from SEM analysis of polychoric correlations are essentially unbiased under assumptions, the $\chi^2$ fit statistic will be substantially inflated and standard errors will be unreliable.

Researchers have attempted to address this problem by applying an asymptotically distribution-free (ADF) approach to SEM analysis of polychoric correlations. The ADF approach uses a discrepancy function of the form

$$F = (\underline{\mathbf{s}} - \underline{\hat{\boldsymbol{\sigma}}})W^{-1}(\underline{\mathbf{s}} - \underline{\hat{\boldsymbol{\sigma}}}),$$

where $\mathbf{s}$ and $\underline{\hat{\boldsymbol{\sigma}}}$ are congruent vectors formed from the nonredundant elements of the empirical ($S$) and model-implied ($\hat{\Sigma}$) covariance/correlation matrix of the observed variables, and $W$ is a consistent estimate of the asymptotic covariance matrix of the elements of $S$. This approach can be applied to SEM analysis with polychoric correlations if one can estimate $W$ for polychoric correlations and then invert that matrix, as indicated. Applied in this way, this approach has been labeled *weighted least squares* (WLS) estimation.

### *Problems With the Weight Matrix*

However, this application also proved to be problematic. The matrix $W$ is large and the fomula is complex. With $k$ observed variables in the structural equation model, $W$ is a square matrix of dimension $k*(k + 1)/2$. This large matrix must be not only estimated but also successfully inverted. Simulation studies have shown this approach to SEM analysis yields stable results only when sample size is very large. Smaller

sample sizes (such as those seen in typical SEM applications) lead to implausible results or to outright estimation failure.

Researchers have described at least two approaches to overcome problems with this WLS approach. One alternative is to estimate the asymptotic covariance matrix empirically through bootstrapping. In a simulation study, this alternative performed well for a model with 15 variables at a sample size of 500. A second alternative, labeled *robust weighted least squares* or *diagonally weighted least squares,* avoids the inversion problem by, in essence, replacing the WLS weight matrix, *W,* with an alternative weight matrix that does not require inversion and whose estimate is stable at substantially lower sample sizes. In simulations, this procedure performed at least as well as the WLS approach and provided acceptable results for small models at sample sizes as low as 200. Violations of the assumption of underlying normality had little effect on the results obtained from either WLS or robust WLS.

### Alternative Approaches for Ordinal Data

Nevertheless, recent research posits a number of alternative approaches for factor analysis of ordinal variables that do not involve the polychoric correlation coefficient. One approach, for example, simultaneously estimates thresholds and factor loadings, with item correlations being inferred from the factor loadings. This approach shifts the focus from responses to individual ordinal variables to patterns of responses across the whole set of observed variables. The approach then maximizes the likelihood of observing the distribution of response patterns actually observed in the data, conditional on the specified factor model. This approach relies on the conditional independence of observed variables given the factor model. Such alternative approaches may likely come to dominate structural equation modeling of ordinal data, limiting the use of the polychoric correlation coefficient for this purpose.

*Edward E. Rigdon*

***See also*** Biased Estimator; Correlation; Estimation; Matrix Algebra; Ordinal Scale; Pearson Product-Moment Correlation Coefficient; Spearman Rank Order Correlation; Structural Equation Modeling

#### Further Readings

Babakus, E., Ferguson, C. E., Jr., & Jöreskog, K. G. (1987). The sensitivity of confirmatory maximum likelihood factor analysis to violations of measurement scale and distributional assumptions. *Journal of Marketing Research, 24,* 222–228.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9,* 466–491.

Jöreskog, K. G. (1994). On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika, 59,* 381–389.

Jöreskog, K. G., & Moustaki, I. (2004). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research, 36,* 347–387.

Lee, S.-Y., Poon, W.-Y., & Bentler, P. M. (1992). Structural equation models with continuous and polytomous variables. *Psychometrika, 57,* 89–105.

Ollson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika, 44,* 443–460.

Yung, Y.-F., & Bentler, P. M. (1994). Bootstrap-corrected ADF test statistics in covariance structure analysis. *British Journal of Mathematical and Statistical Psychology, 57,* 63–84.

## Polynomials

In social-behavioral sciences, the term *polynomial* concerns the use of exponents or special coefficients to assess trends in linear models, with applications in ordinary least squares (OLS) regression, general linear model (GLM) analysis of variance (ANOVA), and other advanced multivariate applications such as hierarchical linear modeling, logistic regression, and generalized estimating equations. In more formal mathematics, polynomials are algebraic expressions that include exponents and have other specific properties. The simplest case of a polynomial is the OLS linear regression equation $Y = A + bX^1$ (in OLS regression, $X^1$ is typically represented as $X$). One of the earliest examples of polynomials to assess trends can be traced to Karl Pearson's 1905 work

addressing "skew correlation," where he used an exponent approach to assess nonlinear association.

The following discussion addresses the use of powered polynomials in OLS regression and orthogonal polynomials derived from special coefficients in both OLS regression and GLM ANOVA (these approaches may be used in other multivariate analyses as well). Both nonorthogonal and orthogonal polynomials are also discussed.

Polynomials can take the form of powered coefficients in OLS linear equations used for trend analysis. The linear equation of $Y = A + b_1X + b_2X^2$ illustrates the use of *powered polynomials,* with $b_1X$ as the linear term (not shown is $X$ taken to the first power), and the squared term $b_2X^2$ producing a quadratic term. When powered polynomials are used for trend analysis, the resulting trends are nonorthogonal, meaning nonindependent. When the trends are created using powered polynomials, the quadratic and higher order terms are built from the linear term (for example, $X$ is used to create $X^2$), and thus are highly correlated. To make the terms orthogonal (i.e., independent), the data require mean-centering as noted by Patricia Cohen, Jacob Cohen, Stephen G. West, and Leona S. Aiken. Otherwise, hierarchical regression is used in assessing the trends to account for nonorthogonality.

Orthogonal polynomials are formed through special coefficients called *coefficients of orthogonal polynomials,* a term used by William Hays, and are provided in most advanced statistics textbooks. They are constructed to be independent of each other. In using orthogonal polynomials for trend analysis, the derived linear and quadratic terms are independent of each other when the group or condition sizes are equal. If group or condition sizes are unequal, the resulting polynomials will not be orthogonal.

The special coefficients for orthogonal polynomials may be used in both OLS regression and GLM ANOVA designs. In regression, the orthogonal polynomial coefficients are used directly as predictor variables for each trend, whereas in ANOVA, the coefficients are used to multiply the group or condition mean values in a fashion similar to those used to assess planned comparisons.

An illustration using orthogonal polynomials in OLS regression is offered below. Assume $Y$ is a behavioral outcome, and $X$ is an ordered grouping variable with three levels representing drug dosage (1 = 0 mg, 2 = 10 mg, 3 = 20 mg). Each group has an $n$ of 30 ($N = 90$). With three dosage levels, a linear and quadratic trend may be fit. Two sets of orthogonal polynomial coefficients taken from David C. Howell are applied to fit these trends: *–1 0 1* for the linear trend; and *1 –2 1* for the quadratic trend. In OLS regression, the resulting regression equation would be $Y = A + bX_{\text{linear}} + bX_{\text{quadratic}}$. The linear trend $X_{\text{linear}}$ is formed by recoding the dosage variable using the orthogonal polynomial coefficients for the linear trend. Those receiving 0 mg of the drug are assigned "–1," those receiving 10 mg are assigned "0," and those with 20 mg of the drug are assigned "1." The same is done for the quadratic trend $X_{\text{quadratic}}$, with those receiving 0 mg assigned "1," those receiving 10 mg assigned "–2," and those receiving 20 mg assigned "1." If this analysis was performed in GLM ANOVA, the same orthogonal polynomial coefficients would be used as planned comparisons to repartition the treatment variance into linear and quadratic components.

*William David Marelich*
*and Elizabeth Grandfield*

***See also*** Analysis of Variance (ANOVA); Multiple Regression; Orthogonal Comparisons; Trend Analysis

**Further Readings**

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Hays, W. L. (1991). *Statistics* (4th ed.). Fort Worth, TX: Harcourt Brace.

Howell, D. C. (2002). *Statistical methods for psychology* (5th ed.). Pacific Grove, CA: Duxbury.

Keppel, G., & Zedeck, S. (1989). *Data analysis for research designs: Analysis of variance and multiple regression/correlation approaches.* New York: Freeman.

Pearson, K. (1905). *On the general theory of skew correlation and non-linear regression.* Drapers' Company Research Memoirs, Biometric Series II. Available at http://books.google.com

# Pooled Variance

The pooled variance estimates the population variance ($\sigma^2$) by aggregating the variances obtained from two or more samples. The pooled variance is widely used in statistical procedures where different samples from one population or samples from different populations provide estimates of the same variance. This entry explains pooled variance, illustrates its calculation and application, and provides cautionary remarks regarding its use.

## Estimate of Population Variance: One Sample Case

In usual research settings, researchers do not know the exact population variance. When there is only one sample from a population, researchers generally use the variance of the sample as an estimate of the population variance. In the case of one sample, the fomula to calculate the sample variance is

$$SD_2 = \frac{\sum (X_i - M)^2}{n - 1},$$

where $X_i$s are the observed scores in the sample, $n$ is the sample size, and $M$ and $SD$ are the sample mean and standard deviation, respectively. This sample variance is an unbiased estimate of the population variance. In other words, the mean of variances of all possible random samples of the *same* size drawn from the population is equal to the population variance.

## Pooled Variance

In many statistical procedures involving multiple groups, there are multiple sample variances that are independent estimates of the same population variance. For example, when samples from the same population are randomly assigned to two or more experimental groups, each group's variance is an independent estimate of the same population variance. In such a condition, the pooled variance is a more precise estimate of the population variance than an estimate based on only one sample's variance. Thus, the variances of all samples are aggregated to obtain an efficient estimate of the population variance.

In the case of $k$ samples whose variances are independent estimates of the same population variance, the fomula to calculate the pooled estimate of the population variance is

$$\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2 + \cdots + (n_k - 1)SD_k^2}{(n_1 - 1) + (n_2 - 1) + \cdots + (n_k - 1)},$$

where $n_1$, $n_2$, . . . , $n_k$ are the sample sizes and $SD_1^2$, $SD_2^2$, . . . , $SD_k^2$ are the sample variances. To obtain the pooled estimate of the population variance, each sample variance is weighted by its degrees of freedom. Thus, pooled variance can be thought of as the weighted average of the sample variances based on their degrees of freedom.

## Homogeneity of Variance

A common application of pooled variance is in statistical procedures where samples are drawn from different populations/subpopulations whose variances are assumed to be similar. In such conditions, each sample's variance is an independent estimate of the same population variance. When variances of samples that belong to different (sub)populations are pooled to yield an estimate of the (sub)population variance, an important assumption is that the (sub)populations have equal variances ($\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$). This assumption is known as the *homogeneity of variance* assumption.

There are statistical tests to evaluate the tenability of the homogeneity of variance assumption, such as Levene's test or Bartlett's test. For these tests, the null hypothesis is that the population variances are equal. It is important to note that the null hypothesis refers to the population parameters. If the null hypothesis is not rejected (i.e., $p_{\text{calculated}}$ value $> \alpha$), the homogeneity of variance assumption holds, and the pooled estimate of the population variance can be used.

## Example

A heuristic example is used to illustrate the calculation and application of pooled variance. Although the number of samples from which we obtain variance statistics will differ, steps used in this example for the pooled-variance calculation can be generalized to other situations. This

example, which compares male and female students' reading scores using a $t$ test, is provided to illuminate applications of pooled variance and processes underlying pooled-variance calculations. Therefore, some details about $t$ tests are explained or elaborated.

The null ($H_0$) and alternative ($H_a$) hypotheses for this $t$ test were stated as follows:

$H_0$ : The population means are the same $(\mu_1 = \mu_2)$.

$H_a$ : The population means are not the same $(\mu_1 \neq \mu_2)$.

To test the $H_0$, two independent samples, 41 female and 51 male students, were randomly chosen from the respective populations. In these samples, the mean reading score and standard deviations for females and males were $M_1 = 100$ $(SD_1 = 50)$ and $M_2 = 80 (SD_2 = 45)$, respectively. Thus, the samples have not only different means but also different variances in their reading achievement scores. Because the extent to which the means are good representatives of the scores in each sample depends on the dispersion of the scores (i.e., a mean better represents all the scores as the $SD$ gets smaller), the dispersions of the scores logically must be taken into account in the comparison of the means.

The $t$ statistic that is used to compare the means of two independent samples is derived using sampling distributions of the differences between two sample means. The fomula for the statistic is

$$t = \frac{(M_1 - M_2) - (\mu_1 - \mu_2)}{SE_{M_1 - M_2}},$$

where $\mu_1 - \mu_2$ is the value stated in the null hypothesis (for our example, $\mu_1 - \mu_2 = 0$) and $SE_{M_1 - M_2}$ is the estimate of the standard error of the difference between sample means $(M_1 - M_2)$.

The standard error of the difference between two sample means in the population $(\sigma_{M_1 - M_2})$ can be obtained using the fomula

$$\sigma_{M_1 - M_2} = \sqrt{\sigma^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)},$$

where $\sigma^2$ is the population variance, and $n_1$ and $n_2$ are sample sizes. It is important to note that to obtain $\sigma_{M_1 - M_2}$, the homogeneity of variance must be assumed (i.e., the variances of the two populations from which the samples are drawn are equal: $\sigma_1^2 = \sigma_2^2 = \sigma^2$). If the homogeneity of variance assumption is not met, $\sigma_{M_1 - M_2}$ can be obtained using adjusted fomulas.

In typical research situations, $\sigma_{M_1 - M_2}$ is rarely known, so an estimate, $SE_{M_1 - M_2}$, is used. If the homogeneity of variance assumption is met, to get the $SE_{M_1 - M_2}$, $\sigma^2$ is replaced by $SD^2_{\text{pooled}}$ in the fomula for $\sigma_{M_1 - M_2}$.

$$SE_{M_1 - M_2} = \sqrt{SD^2_{\text{pooled}} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}$$

In the fomula for $SE_{M_1 - M_2}$, $SD^2_{\text{pooled}}$ is used as an estimate of the population variance. In our example, the sample variances for females and males were 2500 and 2025, respectively. Assuming that the population variances for females and males were the same, each of the sample variances is an independent estimate of the population parameter. Rather than using one of the sample variances as an estimate of the parameter, a more statistically efficient estimate is obtained by pooling the sample variances:

$$SD^2_{\text{pooled}} = \frac{(41 - 1)(50^2) + (51 - 1)(45^2)}{(41 - 1) + (51 - 1)}$$

$$= 2236.1.$$

Thus, the pooled estimate of the population variance, which is a weighted average of the sample variances (i.e., $SD_1^2 = 50$ and $SD_2^2 = 45$ for females and males, respectively) based on their degrees of freedom ($df_1 = 40$ and $df_2 = 50$ for females and males, respectively), is 2236.1.

When the $SD^2_{\text{pooled}}$ value is substituted in the fomula of $SE_{M_1 - M_2}$,

$$SE_{M_1 - M_2} = \sqrt{(2236.1) \left( \frac{1}{41} + \frac{1}{51} \right)} = 9.92$$

and accordingly,

$$t = \frac{(100 - 80) - 0}{9.92} = 2.02.$$

The $t$ statistic is interpreted as a mean difference of 20 in the sample and is 2.02 standard deviations away from the hypothesized population mean difference, which is 0. Using a commonly available $t$ statistic distribution, the $p$ value for our two-tailed hypothesis test is approximately .046. At the .05 level of significance, we reject the $H_0$, which stated that the reading scores for females and males in the population were the same.

In addition to $t$ tests, pooled variances are commonly used as part of analyses of variance (ANOVAs). For example, when conducting an ANOVA to compare differences between students' mean reading scores by ethnicity, each ethnicity is a subgroup of the student population. To test differences of mean reading scores between ethnicities, an important assumption in ANOVA is that the variances between dependent variable scores in each ethnicity in the population are equal. Thus, a pooled estimate of the population variance is obtained by aggregating the sample variances from each ethnicity. ANOVA is somewhat robust to the violation of homogeneity of variance assumption, if group sizes are equal, so in many cases, a pooled variance is acceptable even if the variances in the populations are not considered exactly equal.

### Careful Use of Pooled Variances

Introductory statistics courses typically introduce parametric methods, which have more assumptions about the data than nonparametric methods. For example, parametric assumptions for a $t$ test for comparing two means are that (a) data have been randomly chosen from the population, (b) the population data are normally distributed, and (c) the variances in the two populations are equal (i.e., homogeneity of variance). When these assumptions are met, parametric methods produce accurate results and the $t$ statistic would be computed with the pooled variance and evaluated with $n_1 + n_2 - 2$ degrees of freedom. However, as the variances in the two samples diverge, the better choice would be to use the unpooled variance along with the appropriate degrees of freedom.

In the case of the $t$ test, unpooled variances would be used if the $t$ test assumptions were not met. The unpooled variance equals the sum of the variances divided by the sample size for each of the two samples. When sample sizes of the two groups are equal, the pooled and unpooled variances are equivalent. The impact of using the pooled or unpooled variance is affected by small sample sizes and whether or not the larger variance is associated with the smaller or larger sample size.

When parametric assumptions are met and sample sizes are not equal, the absolute value of the $t$ statistic will be slightly higher when the unpooled variance is used. The higher $t$ statistic corresponds to a higher likelihood of rejecting the null hypothesis and an inflated risk of committing a Type I error (i.e., rejecting a true hypothesis). There has been some discussion of using the unpooled variances more often, and the interested reader can consult the Further Readings section to find out more about that discussion.

*Linda Reichwein Zientek and Z. Ebrar Yetkiner*

***See also*** Homogeneity of Variance; Standard Error of Estimate; $t$ Test, Independent Samples; Variance

**Further Readings**

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2003). *Applied statistics for the behavioral sciences* (5th ed.). Boston: Houghton Mifflin.

Julios, S. A. (2005). Why do we use pooled variance analysis of variance? *Pharmaceutical Statistics, 4,* 3–5.

Moser, B. K., & Stevens, G. R. (1992). Homogeneity of variance in the two-sample means test. *The American Statistician, 46,* 19–21.

Ruxton, G. D. (2006). The unequal variance $t$ test is an underused alternative to Student's $t$ test and the Mann–Whitney U test. *Behavioral Ecology, 17,* 688–690.

Sheskin, D. J. (2004). *Handbook of parametric and nonparametric statistical procedures* (3rd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Thompson, B. (2006). *Foundations of behavioral sciences: An insight-based approach*. New York: Guilford.

## Population

With respect to research design and statistical analysis, a *population* is the entire collection of entities one seeks to understand or, more formally, about which one seeks to draw an inference.

Consequently, defining clearly the population of interest is a fundamental component of research design because the way in which the population is defined dictates the scope of the inferences resulting from the research effort.

When a population is small, it can be censused, meaning that the attribute of interest is measured for every member of the population and is therefore known with certainty. More often, however, a census is unrealistic because the population is not finite or because financial or logistical constraints override collection of complete information. In some cases, simply redefining the population may permit a census, perhaps by reducing the size or extent of the population of interest. When a census is still not feasible or when complete information about a population is unnecessary, a subset of the defined population called a *sample* is selected and measured. Attributes of the sample are used to make inferences about the entire population from which the sample was drawn. Statistical tools make use of these inductive inferences, using specific cases (the samples) to generalize or make inferences about the whole (the population). Specifically, a sample statistic provides an estimate of the population parameter, the true value of the attribute of interest. As George Snedecor and William Cochran wrote, it is the sample we observe, but the population we wish to know.

Such inductive inferences will be valid when the sample is drawn using an approach that ensures each unit of the population has a chance to be selected as part of the sample, such as a probabilistic sampling scheme, and inferences are restricted to the entities of the population that were available for sampling. For example, if data on test scores were collected from a random sample of undergraduate students enrolled in a specific calculus course at a single university, inferences apply only to that course and university and not to all undergraduates, all math courses, or all universities. If a researcher was interested in drawing general conclusions about current test scores in undergraduate calculus courses throughout the United States, undergraduate students currently enrolled in all calculus courses at universities across the United States must be available for selection as part of the random sample. This example emphasizes the importance of defining the population of interest explicitly when designing a sampling scheme because of its overriding importance on inference.

*Andrea R. Litt*

***See also*** Error; Inference: Deductive and Inductive; Parameters; Precision; Research Design Principles; Sample; Sampling; Statistic

**Further Readings**

Scheaffer, R. L., Mendenhall, W., III, & Ott, L. (2006). *Elementary survey sampling* (6th ed.). Belmont, CA: Wadsworth.

Snedecor, G. W., & Cochran, W. G. (1967). *Statistical methods* (6th ed.). Ames: Iowa University Press.

Thompson, S. K. (2002). *Sampling* (2nd ed.). New York: Wiley.

# Positivism

Positivism is the name of a philosophical doctrine created in France by Auguste Comte. As a term of research in human sciences, *positivism* has come to be closely associated with the idea of fact-based investigation, being a pejorative variation of *empiricism*. This entry discusses the various definitions and applications of positivism.

## Definitions

When *positivism* as a term is used in the history of historiography, it tends to refer to the work of scholars in the Third Republic in France; *positivistic historians* in this context means scholars who produced historical research that followed a rigorous method of systematic investigation of sources (what historians of the Annales school pejoratively called *histoire historisante*). In social sciences, *positivism* has become a synonym for a rejection of metaphysics and a heightened concern for observable and measurable phenomena in the natural and social world. Likewise, in language studies, the term *positivism* has come to mean a failed attempt to compare and transplant methods of physical sciences to the realm of human language-based communication. Until recently, in all these cases, when the word *positivism* is used, it was normally used in a negative sense to allow for

criticism of the approach or method in question (Karl Popper having been one of the main opponents of Comte's philosophy of science in the 20th century).

The word *positivism* was incorporated into the vocabulary of human sciences, and it currently has different meanings for modern scholars. Georg Iggers claimed that positivism means a belief that historical facts are self-explanatory and do not require mediation of hypothesis, interpretation, or operational definitions. Wilson Simon claimed that only Comte's positivism could truly be called by that name, as it was based on a personal conviction that human behavior could be examined using the same laws of physical and natural sciences. Donald Charlton, Norman Cantor, and Richard Schneider promoted the idea of positivism replaced with a "scientific nominalism," meaning textual criticism and hostility to hidden meaning and interpretation was a more apt term than positivism. For William Keylor, positivism means a theory of fact-oriented science.

## Applications

Any initial definition of the term *positivism,* however, should relate it to Comte's work. Comte's oeuvre sought to establish an understanding of how humanity had evolved from the first stage of its development (theological) to the second (metaphysical) and would further progress until it finally reached its most perfect state (the positive stage). Positivism—or *science positive*—is Comte's ambitious attempt to synthesize all the conflicting philosophies of his time through a new philosophy of history capable of explaining human experience as a function of the development of the human mind. Through this and through the epistemological assumption that only through direct observation can knowledge be attained, Comte envisioned human history in somewhat utopian terms as culminating in a flawlessly well-integrated society. In order to capture all of human experience (historical, social, political, and intellectual), Comte founded the science of sociology.

Combining "social positivism" (the science that studied the human race, focusing on the social relationships between individuals); the "religion of humanity" (the ethical, nontheistic religion based on the collective and the individual aspects of humankind); and a theory of knowledge (based on the assumption that humans evolve from the theological stage into the metaphysical one and then to the positive one, not yet fully achieved), positivism forms the basis of modern science, at which stage science gains supremacy over philosophy (or metaphysics). In this context, positivism can be seen as an attempt to incorporate and improve the pre-19th-century theories of knowledge, which based some or all of their arguments on a priori knowledge and relied upon metaphysical, religious, or moralistic assumptions, not simply on material reality. The purpose of Comte's work was to offer a satisfactory explanation for the crisis in French politics and society, which had started with the French Revolution in 1789, and to suggest that by understanding social phenomena, one could understand political occurrences and, as a result, start the process of social reconstruction that would eventually eliminate politics from society.

Positivism entered the debates of politicians, academics, and others in the 1850s, becoming dominant, at least in France, by the 1860s. The positivism to which these groups referred was not always Comtean positivism. Comte's positivism as applied to social sciences (he actually coined the term *sociology,* capturing his life's goal of understanding human experience) was a system based on an understanding of the world and humankind founded on the principle of an analogy between the elements of the human world (complex phenomena of language, social life, and human thought) and those of the natural world (simple phenomena of inorganic and organic nature). In Comte's system, the analogy between humans and nature allowed the scientist to apply the methods of the positive sciences (biology, physics, and chemistry) to the humanities. In particular, the use of Comte's law of three stages of humanity, whereby societies could evolve from one less developed stage to another, more developed stage, was an essential part of 19th-century philosophy of history, which, although incorporating the works of some like Condorcet (Comte's spiritual father, from whom he learned the idea of organic development of human history in 10 stages), largely rejected the Whiggish/Enlightenment-based idea of history as a linear progression as opposed to a series of developmental stages. The development-based approach perfected by Comte clearly influenced

many 19th-century thinkers, including John Stuart Mill and Karl Marx. For both Comte and Marx, the political system reflected civil order, this in turn reflecting the state of civilization.

An element of Comte's philosophy that was appropriated by scholars in the second half of the 19th century was the claim that education was the basis for social development because it was capable of regenerating society. Comte's pedagogical claims are varied, mostly opposing the humanist-type education (which valued rhetoric and the power of language) and focusing on clarity, objectivity, and conciseness. In Comtean philosophy, the study of facts is essential because facts are morally charged. Consequently, the investigation of facts has the capacity of social transformation. The need is, therefore, to ensure that facts were learned the appropriate way and according to pupils' developmental stage. Comte posited an analogy between the mental development of individuals and that of humanity, so that the individual's education progresses through the same intellectual/emotional stages as humanity has progressed through history, until it finally reaches the positive stage. This fomulation and its uses by educators were particularly important because of its pedagogical implications and applications (exemplary cases of the development of positivistic-based pedagogy were developed in India, Brazil, Mexico, Guatemala, and Argentina, as well as France itself), extending in time, in some cases, to the post-World War II period.

Aside from sociology and pedagogy, positivism had a short but powerful application to French politics. Comtean positivism was, by nature, authoritarian, hierarchic, and antidemocratic, based on a reactionary view that was intended as a means to guarantee a nonviolent political environment and the realization of social perfection through peaceful means. Comte acknowledged himself as a follower of Joseph de Maistre and Louis de Bonald, both of whom were counter-Revolutionary thinkers who claimed that there was no need for violent revolution to change society. This view generally has been accepted, although some revisionist work has sought to demonstrate that many of the followers of Comte aimed to use positivism in their creation of liberal and democratic institutions. It is undeniable, however, that the aims of Comte himself were far from liberal and democratic. Order and progress were the tenets of the project that would eventually implement a positive society. This was accomplished via a series of highly complex scientific and emotional/religious processes whereby wisemen were responsible for educating the lower classes and establishing a republic of knowledge.

In France, positivism came to be associated with republicanism, in particular during the latter years of the Second Empire and the dawn of the Third Republic. Sudhir Hazareesingh suggests that the republican interpretation of Comte's positivism was offered by Émile Littré, a disciple of Comte and an early enthusiast of his master's version of positivism. It was through him that the Comtean doctrine was liberalized and republicanized, thus losing its affiliation to its creator and becoming a philosophy in and of itself; some of the European variations of positivism are clearly Littrean. Littré wrote *Conservation, Revolution, Positivisme* in 1852 to recover the idea that positivism could have a socially and politically transformative character as long as it became clear that under a centralized government such as that which the Second Empire was shaping up to be, there could be no unified science and no hope for progress. What Littré passed on to republicans like Third Republic figures Léon Gambetta and Jules Ferry was the idea that one can create order from disorder, as Comte had advocated, but scientific progress was possible only under a republic. In order for there to be a republic, there should be democracy, and in order for there to be democracy, socialized (mass) education was necessary.

In Latin America, however, the positivism that influenced governments in the late 19th century was hardly the democratic-Littrean variation, but rather the authoritarian and conservative Comtean positivism, as is clear in the cases of Brazil, Mexico, Argentina, and Guatemala. In any event, the political application of positivism sought to legitimize a strong state and to focus on mass education as a means to guarantee adherence to the status quo, as Leopoldo Zea claimed.

The influence of positivism in politics and pedagogy was indeed felt for nearly a century in Europe and abroad. As a philosophy of science and a research methodology, however, the supremacy of positivism was short-lived. By the end of the 19th century, positivism, as a methodology for

human sciences, had fallen into disrepute. Linguists and philologists became increasingly suspicious of the mixture of physical and human sciences as regards their methodology. Sociologists felt the same, and even historians started disclaiming theories that linked the process of knowing rocks as described by geologists to the process of knowing human beings as sought by historians, as Charles-Olivier Carbonell has noted. Nonetheless, there were attempts to recover Comtean positivism as a methodology for human sciences in the late 19th century, especially as the process of secularization of French society and politics was under way. Denying the current state of institutionalized religion, positivism's erasure of free will, and Christian providentialism in favor of a deterministic and accident-based approach (as Raymon Aron suggested), at least as a label, positivism still portrayed itself as a methodological guarantor of the scientific rigor of historical studies and had fulfilled its role.

Debates about positivism in historical studies and human sciences in general had one effect: In the desire to show their usefulness for society, human sciences (and their scholars) attempted to help themselves by establishing a link between democracy, liberalism, and liberal arts. History, which was once seen as a mildly conservative discipline in the sense of pointing to alternatives to drastic social change, became the herald of a new order. Sociology, too, by reinterpreting positivism via the work of Émile Durkheim, established itself as a discipline (also of slightly conservative tendencies) in Paris and then other European nations thanks to positivism.

The sound bites of positivism are varied and can be found in numerous sources that contend, dismiss, and detract positivism and the aims of its conceptual creator: the need for objectivity and an empirical approach to facts; the lack of need to look for final causes for historical facts; an understanding that facts are morally charged and, as a result, capable of transforming politics and society; the predictive ability of science in terms of pointing to patterns of human behavior; and the need to use direct observation in order to accrue knowledge (in the case of past knowledge, empirical use of documentary sources, mostly literary), in addition to the moral need to educate the politically enabled populace. They point to the variety of ways in which positivism was employed (and is still employed) in politics, epistemology, and moral philosophy.

*Isabel DiVanna*

***See also*** Cause and Effect; Critical Theory; *Logic of Scientific Discovery, The;* Scientific Method; Threats to Validity

**Further Readings**

Arbousse-Bastide, P. (1957). *La doctrine de l'éducation universelle dans la philosophie d'Auguste Comte. Principe d'unité systématique et fondement de l'organisation spirituelle du monde* [The doctrine of universal education in the philosophy of Auguste Comte. Principle of systematic unity and foundation of the organization's spiritual world]. 2 vols. Paris: PUF.

Cantor, N., & Schneider, R. (1967). *How to study history.* New York: Harlan Davidson.

Charlton, D. G. (1959). *Positivist thought in France during the Second Empire, 1852–1870.* Oxford, UK: Oxford University Press.

Keylor, W. R. (1975). *Academy and community: The foundation of the French historical profession.* Cambridge, MA: Harvard University Press.

Le(s) positivisme(s). (1978). Special issue. *Revue Romantisme, 21–22.*

Simon, W. (1963). *European positivism in the nineteenth century: An essay in intellectual history.* Ithaca, NY: Cornell University Press.

# Post Hoc Analysis

Post hoc analysis applies to tests of differences among sample statistics when the specific hypothesis to be tested has been suggested by the values of the statistics themselves. Perhaps the most common statistic to be tested is the sample mean in experiments involving three or more means. In such cases, it is most common to begin with the testing of an analysis of variance (ANOVA) $F$ test. A nonsignificant $F$ test implies that the full null hypothesis of the equality of all population means is plausible. Consequently, no additional testing would be considered.

Suppose a significant ANOVA $F$ test is found in an experiment involving $k = 4$ means. If the last

two means are greater than the first two means, a researcher might want to know if that difference would also be found in comparing the corresponding population means. That researcher would be interested in testing the hypothesis

$$H_0 : \frac{\mu_1 + \mu_2}{2} = \frac{\mu_3 + \mu_4}{2}.$$

Equivalent hypotheses would be

$$H_0 : \mu_1 + \mu_2 - \mu_3 - \mu_4 = 0$$

or

$$H_0 : \psi = 0,$$

where $\psi$ represents the ordered coefficients (1, 1, –1, –1) applied to the corresponding four population means. In general, any contrast among four means could be expressed by the four ordered coefficients ($c_1$, $c_2$, $c_3$, $c_4$).

To test the hypothesis using the four sample means, $\psi$ is estimated by applying the four coefficients to the corresponding sample means to calculate $Y$ as

$$Y = c_1 M_1 + c_2 M_2 + c_3 M_3 + c_4 M_4.$$

For equal sample sizes with common value, $N$, a corresponding sum of squares is obtained by

$$SS_Y = NY_2 / c_i^2.$$

For unequal sample sizes, $N_1, N_2, \ldots, N_k$, the fomula is

$$SS_Y = \frac{Y^2}{\sum_{i=1}^{i=k} \frac{c_i^2}{N_i}}.$$

The contrast has $df = 1$, so dividing by l (i.e., keeping the same value) changes the $SS_Y$ to a mean square, $MS_Y$. If the contrast had been chosen without first examining the sample means (especially if it was chosen before the experiment was performed), then it would be an a priori contrast rather than a post hoc contrast. In that case, an $F$ test for the contrast, $F_Y$, is calculated by

$$F_Y = \frac{MS_Y}{MS_{\text{error}}},$$

where $MS_{\text{error}}$ is the denominator of the ANOVA $F$ test. The critical value would be obtained from an $F$ distribution with 1 and $df_{\text{error}}$ degrees of freedom.

For post hoc analysis, Henry Scheffé provided a similar method for testing the contrast hypothesis. For a contrast among $k$ means, the $MS_Y$ is

$$MS_Y = \frac{SS_Y}{k - 1}.$$

The critical value is obtained from an $F$ distribution with $k$ – 1 and $df_{\text{error}}$ degrees of freedom. If the critical value is determined with a significance level, $\alpha$ the probability of one or more Type I errors will be limited to $\alpha$ no matter how many contrasts are evaluated as post hoc contrasts by the Scheffé procedure. In fact, Scheffé even proved that if the ANOVA $F$ test is significant, then at least one post hoc contrast must be significant by the Scheffé procedure.

Scheffé's procedure can be used for testing hypotheses about any number of means in a contrast. However, pairwise testing of only two means at a time can be done with more powerful procedures. John Tukey proposed a single, critical difference, CD, for all pairs of means in a group of $k$ means each with sample size $N$. That value is

$$CD = q_{1-\alpha}(k, df_{\text{error}}) \sqrt{\frac{MS_{\text{error}}}{N}},$$

where $q_{1-\alpha}(k, df_{\text{error}})$ is the $100(1 - \alpha)$ percentage point of the Studentized range distribution with parameters $k$ and $df_{\text{error}}$. Tukey's procedure limits the probability of one or more Type I errors to $\alpha$ even when testing all $k(k - 1)/2$ pairs of $k$ means and even if the ANOVA $F$ test is *not* significant. However, it is almost a universal practice to apply Tukey's procedure only after a significant ANOVA $F$ test.

With unequal sample sizes $N_i$ and $N_j$, Tukey's procedure is modified to become the Tukey–Kramer procedure where CD is given by

$$\text{CD} = q_{1-\alpha}(k, df_{\text{error}}) \sqrt{\frac{MS_{\text{error}}}{2} \left[ \frac{1}{N_i} + \frac{1}{N_j} \right]}.$$

The Tukey–Kramer procedure also limits the probability of one or more Type I errors to $\alpha$ even

when testing all $k(k-1)/2$ pairs of $k$ means and even if the ANOVA $F$ test is not significant. However, it is also almost a universal practice to apply the Tukey–Kramer procedure only after a significant ANOVA $F$ test.

A. J. Hayter proposed a modification of Fisher's least significant difference procedure that can be considered a powerful variation on Tukey's procedure. For equal $N$, the Hayter–Fisher CD becomes

$$\text{CD} = q_{1-\alpha}(k-1, df_{\text{error}})\sqrt{\frac{MS_{\text{error}}}{N}}.$$

The only change from Tukey's CD fomula is to replace the parameter $k$ with $k-1$. This makes the CD value for the Hayter–Fisher procedure always smaller than the CD for Tukey's procedure. However, the Hayter–Fisher procedure always requires a significant ANOVA $F$ test before applying the CD to pairwise testing. For pairwise testing of exactly three means, each based on the same size, $N$, and following a significant ANOVA $F$ test, there is no procedure more powerful than the Hayter–Fisher when the usual ANOVA assumptions are satisfied.

For unequal $N$, the Hayter–Fisher uses the fomula

$$\text{CD} = q_{1-\alpha}(k-1, df_{\text{error}})\sqrt{\frac{MS_{\text{error}}}{2}\left[\frac{1}{N_i}+\frac{1}{N_j}\right]}.$$

Again, the only change from the Tukey–Kramer is replacing $k$ with $k-1$.

For $k > 3$ means, a procedure more powerful than the Hayter–Fisher is the Peritz procedure. The procedure involves following a significant ANOVA $F$ test with additional $F$ tests for all possible tests of null conditions of partitions of the $k$ means. The procedure is illustrated in the example below.

The statistical literature provides a large number of alternative procedures for post hoc testing. However, the few procedures presented here can be used in many cases. To avoid the ANOVA assumption of equal population variances, the Peritz $F$ tests can be replaced by the Brown and Forsythe procedure as modified by D. V. Mehrotra.

## Illustrative Example

Alizah Z. Brozgold, Joan C. Borod, Candace C. Martin, Lawrence H. Pick, Murray Alpert, and Joan Welkowitz reported a study of marital status for several psychiatric groups and normal controls with high scores indicating *worse* marital status. The summary statistics for four groups were as follows:

| *Group 1* | *Group 2* | *Group 3* | *Group 4* |
|---|---|---|---|
| *Right Brain Damaged* | *Normal Controls* | *Unipolar Depressives* | *Schizophrenics* |
| *RBD* | *NC* | *UD* | *SZ* |
| $M_1 = 1.22$ | $M_2 = 1.50$ | $M_3 = 3.00$ | $M_4 = 4.61$ |
| $s_1 = 1.93$ | $s_2 = 1.79$ | $s_3 = 1.91$ | $s_4 = 1.69$ |
| $N_1 = 19$ | $N_2 = 21$ | $N_3 = 12$ | $N_4 = 20$ |

| *Source of Var.* | *SS* | *df* | *MS* | *F* | *p Value* |
|---|---|---|---|---|---|
| Treatment | 144.076 | 3 | 48.026 | 14.48* | $< .0001$ |
| Within Groups | 225.556 | 68 | 3.317 | | |
| Total | 3369.633 | 71 | | | |

*$p < .05$. Reject $H_0$ because $14.48 > 2.75 = F_{.95}(3,65) > F_{.95}(3,68) = \text{CV}$.

Table 1 presents the CD values for the Scheffé, Tukey–Kramer, and Hayter–Fisher procedures. As expected, the Scheffé is least powerful for pairwise testing, showing only two differences as significant. In this case, the Right Brain Damaged have significantly better marital status than the

**Table 1** Six Pairwise Differences Tested for Significance by Three Procedures

| *Pair* | *Group i* | *Group j* | *Diff.* | *Scheffé-CD* | *TK-CD* | *HF-CD* |
|---|---|---|---|---|---|---|
| 1 | 1.22 | 1.50 | .25 | 1.65 | 1.50 | 1.37 |
| 2 | | 3.00 | 1.78 | 1.93 | 1.75* | 1.60* |
| 3 | | 4.61 | 3.39 | 1.67* | 1.52* | 1.35* |
| 4 | 1.50 | 3.00 | 1.50 | 1.89 | 1.71 | 1.57 |
| 5 | | 4.61 | 3.11 | 1.63* | 1.48* | 1.35* |
| 6 | 3.00 | 4.61 | 1.61 | 1.91 | 1.73 | 1.58* |

*Source:* Brozgold, A. Z., Borod, J. C., Martin, C. C., Pick, L. H., Alpert, M., & Welkowitz, J. (1998). Social functioning and facial emotional expression in neurological and psychiatric disorders. *Applied Neuropsychology, 5,* 15–23.

*Note:* *Difference significant at $\alpha = .05$.

Schizophrenics. The Tukey–Kramer is more powerful, showing three differences as significant. That is, the Tukey–Kramer also finds significantly better marital status for the Right Brain Damaged than the Unipolar Depressives. Finally, the most powerful of the three is the Hayter–Fisher, with four significant differences. The Hayter–Fisher finds the Unipolar Depressives to have significantly better marital status than the Schizophrenics.

Table 2 presents the $F$ tests required by the Peritz procedure that are needed to determine the significance of the difference between Means 2 and 3 (i.e., $M_2 = 1.50$ and $M_3 = 3.00$). This pair of means was not significantly different in Table 1. The first partition is actually no partition at all because it includes all four means. The $F$ test is the overall ANOVA $F$ test and is significant as noted above. The second partition includes the first three means in one group, and Mean 4 is a second group. The hypothesis is that the first three population means are equal, and the $F = 3.83$ is significant at $\alpha = .05$ so the hypothesis is rejected. The third partition assigns $M_1$ and $M_4$ to one group and $M_2$ and $M_3$ to a second group. Both groups are tested for a significant difference at $\alpha = .0253$. Only one need be significant for the Peritz procedure to find $M_2$ and $M_3$ to differ significantly. The fourth partition assigns $M_1$ to one group and $M_2$, $M_3$, and $M_4$ to a second group. The hypothesis is that the last three population means are equal. This hypothesis is also significant at $\alpha = .05$. Finally, the fifth partition assigns $M_1$ to one group, $M_2$ and $M_3$ to a second group, and $M_4$ to a third group. This partition tests the simple hypothesis that the second and third population means are identical. This test is significant at $\alpha = .05$. The five partitions shown in Table 2 include all partitions in which $M_2$ and $M_3$ are included together in the same group. A significant test must be found for all five partitions.

The additional significant pair found by the Peritz procedure illustrates its greater power than the other three procedures. The Normal Controls are shown to have significantly better marital status than the Unipolar Depressives.

The entire Peritz procedure requires the same type of testing shown in Table 2 applied to all six pairs of means. The Peritz procedure identifies five of the six pairs to be significantly different. The Peritz procedure will generally (but not always) be more powerful than the Hayter–Fisher procedure for $k \geq 4$. Clearly, this procedure requires a computer program to be practical.

*Philip H. Ramsey*

***See also*** Mean Comparisons; Multiple Comparison Tests; Pairwise Comparisons

**Table 2** Pairwise Testing With the Peritz Procedure Applied to Means $M_2 = 1.50$ and $M_3 = 3.00$

| | | *M1* | *M2* | *M3* | *M4* | | |
|---|---|---|---|---|---|---|---|
| *Partition* | *Hyp.* | *1.22* | *1.50* | *3.00* | *4.61* | *F* | $F_{CV}$ |
| 1 | | 1 | 1 | 1 | 1 | | |
| | 1 | 1 | 1 | 1 | 1 | 14.48 | 2.74 |
| 2 | | 1 | 1 | 1 | 2 | | |
| | 1 | 1 | 1 | 1 | | 3.83 | 3.13 |
| 3 | | 1 | 2 | 2 | 1 | | |
| | 1 | 1 | | | 1 | 33.76 | 5.23 |
| | 2 | | 2 | 2 | | 5.18 | 5.23 |
| 4 | | 1 | 2 | 2 | 2 | | |
| | 1 | | 2 | 2 | 2 | 14.94 | 3.13 |
| 5 | | 1 | 2 | 2 | 3 | | |
| | 1 | | 2 | 2 | | 5.18 | 3.98 |

## Further Readings

Brown, M. B., & Forsythe, A. B. (1974). The small sample behavior of some statistics which test the equality of several means. *Technometrics, 16,* 129–132.

Brozgold, A. Z., Borod, J. C., Martin, C. C., Pick, L. H., Alpert, M., & Welkowitz, J. (1998). Social functioning and facial emotional expression in neurological and psychiatric disorders. *Applied Neuropsychology, 5,* 15–23.

Hayter, A. J. (1984). A proof of the conjecture that the Tukey–Kramer multiple comparisons procedure is conservative. *Annals of Statistics, 12,* 61–75.

Hayter, A. J. (1986). The maximum familywise error rate of Fisher's least significant difference test. *Journal of the American Statistical Association, 81,* 1000–1004.

Mehrotra, D. V. (1997). Improving the Brown-Forsythe solution to the generalized Behrens-Fisher problem. *Communications in Statistics, 26,* 1139–1145.

Ramsey, P. H. (2002). Comparison of closed procedures for pairwise testing of means. *Psychological Methods, 7,* 504–523.

Scheffé, H. E. (1959). *The analysis of variance*. New York: Wiley.
Tukey, J. W. (1953). *The problem of multiple comparisons*. Unpublished manuscript, Princeton University, Princeton, NJ.

# Post Hoc Comparisons

The *F* test used in analysis of variance (ANOVA) is called an *omnibus* test because it can detect only the presence or the absence of a global effect of the independent variable on the dependent variable. However, in general, we want to draw specific conclusions from the results of an experiment. Specific conclusions are derived from focused comparisons, which are, mostly, implemented as contrasts between experimental conditions. When these comparisons are decided after the data are collected, they are called *post hoc* or *a posteriori analyses*. These comparisons are performed after an ANOVA has been performed on the experimental data of interest. In the ANOVA framework, post hoc analyses take two general forms: (a) comparisons that involve all possible contrasts and (b) comparisons that are restricted to comparing pairs of means (called *pairwise comparisons*).

## Notations

The experimental design is a one-factor ANOVA. The total number of observations is denoted *N*, with *S* denoting the number of observations per group. The number of groups is denoted *A*, a given group is labeled with the letter *a*, the group means are denoted $M_{a+}$, and the grand mean is denoted $M_{++}$. A contrast is denoted $\psi_a$, the contrast coefficients (contrast weights) are denoted $C_a$. The $\alpha$-level per comparison is denoted $\alpha[PC]$ and the $\alpha$-level per family of comparisons is denoted $\alpha[PF]$.

## Planned Versus Post Hoc Comparisons

Planned (or a priori) comparisons are selected before running the experiment. In general, they correspond to the research hypotheses. Because these comparisons are planned, they are usually few in number. In order to avoid an inflation of the Type I error (i.e., declaring an effect significant when it is not), the *p* values of these comparisons are corrected with the standard Bonferroni or Šidàk approaches.

In contrast, post hoc (or a posteriori) comparisons are decided after the experiment has been run and analyzed. The aim of post hoc comparisons is to make sure that (unexpected) patterns seen in the results are reliable. This implies that the actual family of comparisons consists of all possible comparisons, even if they are not explicitly tested.

## Contrast

A contrast is a prediction precise enough to be translated into a set of numbers called contrast coefficients or contrast weights (denoted $C_a$) that express this prediction. For convenience, contrast coefficients have a mean equal to zero. The correlation between the contrast coefficients and the values of the group means quantifies the similarity between the prediction and the results.

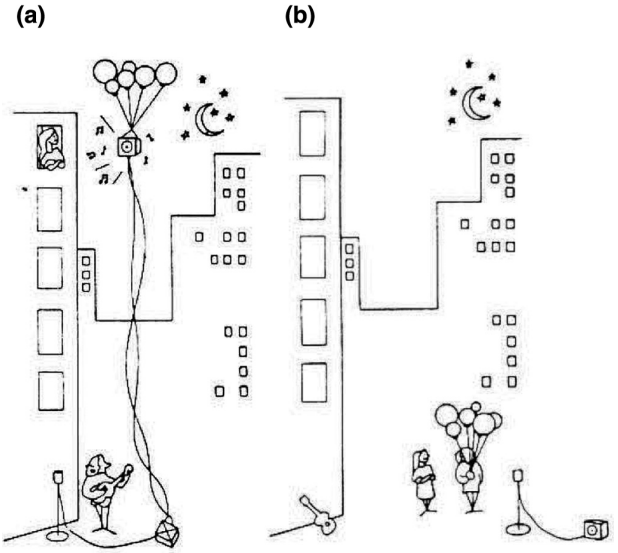



**Figure 1** (a) Appropriate Context (b) Partial Context for Bransford and Johnson's (1972) Experiment

*Source:* Reprinted from *Journal of Verbal Learning and Verbal Behavior, 11,* Bransford, J.D. Contextual prerequisites for understanding: Some investigations of comprehension and recall," 717–726, Copyright 1972, with permission from Elsevier.

All contrasts are evaluated using the same general procedure. A contrast is formalized as a set of contrast coefficients that represents the predicted pattern of experimental results. For example, if we have four groups and we predict that the first group should have better performance than the other three groups, and that these remaining three groups are equivalent, we get the following contrast:

$$\begin{array}{rccccc} & C_1 & C_2 & C_3 & C_4 & \text{Mean} \\ \psi = & 3 & -1 & -1 & -1 & 0 \end{array} \quad (1)$$

When the data have been collected and the experimental means and mean squares have been computed, the next step is to evaluate if the contrast correlates significantly with the mean values. In order to do so, a specific $F$ ratio (denoted $F_\psi$) is computed. Finally, the probability associated with $F_\psi$ is evaluated. If this probability is small enough, the contrast is considered significant.

## Example

This example is a fictitious replication of Bransford and Johnson's 1972 study examining the effect of context on memory. In this study, Bransford and Johnson read the following paragraph to their participants:

> If the balloons popped, the sound would not be able to carry since everything would be too far away from the correct floor. A closed window would also prevent the sound from carrying since most buildings tend to be well insulated. Since the whole operation depends on a steady flow of electricity, a break in the middle of the wire would also cause problems. Of course the fellow could shout, but the human voice is not loud enough to carry that far. An additional problem is that a string could break on the instrument. Then there could be no accompaniment to the message. It is clear that the best situation would involve less distance. Then there would be fewer potential problems. With face to face contact, the least number of things could go wrong.

To show the importance of context on memory for texts, the authors used the following four experimental conditions:

1. *No context*: Participants listened to the passage and tried to remember it.
2. *Appropriate context before*: Participants were provided with an appropriate context (see Figure 1a) in the form of a picture and then listened to the passage.
3. *Appropriate context after*: Participants first listened to the passage and then were provided with an appropriate context (see Figure 1a) in the form of a picture.
4. *Partial context*: Participants were provided with a context that did not allow them to make sense of the text when they listened to it (see Figure 1b).

The dependent variable is the number of "ideas" recalled. The data are shown in Table 1, and the results of the ANOVA are shown in Table 2.

**Table 1** Number of Ideas Recalled for 20 Subjects in the Fictitious Replication of Bransford and Johnson (1972)

| | *Context Before* | *Partial Context* | *Context After* | *No Context* |
|---|---|---|---|---|
| | 5 | 5 | 2 | 3 |
| | 9 | 4 | 4 | 3 |
| | 8 | 3 | 5 | 2 |
| | 4 | 5 | 4 | 4 |
| | 9 | 4 | 1 | 3 |
| $\Sigma$ | 35 | 21 | 16 | 15 |
| $M_{a+}$ | 7 | 4.2 | 3.2 | 3 |
| $M_{++=}4.35$ | | | | |
| $S = 5$ | | | | |

*Note:* The experimental conditions are ordered from the smallest to the largest mean.

**Table 2** ANOVA Results for the Bransford and Johnson (1972) Experiment

| *Source* | *df* | *SS* | *MS* | *F* | *p(F)* |
|---|---|---|---|---|---|
| Between | 3 | 50.90 | 16.97 | 7.22 | .00288 |
| Error | 16 | 37.60 | 2.35 | | |
| Total | 19 | 88.50 | | | |

*Source:* Adapted from Bransford and Johnson (1972).

## Post Hoc Comparisons

### *Scheffé Test*

When a post hoc comparison is performed, the family of comparisons consists of all possible comparisons. This number is, in general, very large, even when dealing with a small number of experimental conditions, and this large number practically precludes using a Bonferroni approach to correct for multiple testings because this large number of comparisons would make the correction much too conservative. The Scheffé test is also conservative, but less than the Bonferonni approach. The idea behind the Scheffé test begins with the omnibus $F$ testing the null hypothesis that all population means are equal, and this null hypothesis in turn implies that all possible contrasts are also zero. Note that testing the contrast with the largest sum is equivalent to testing all possible contrasts at once (because a failure to reject the null hypothesis with the largest contrast implies a failure to reject the null hypothesis for any smaller contrast).

Suppose that Bransford and Johnson wanted to test the following four contrasts *after* having collected their data: (a) no context versus all other conditions, (b) context after versus no context, (c) context before versus all other conditions, and (d) context before versus partial context. The contrast weights (i.e., $C_a$s) are shown in Table 3.

The $F_\psi$ ratio for the maximum contrast $\psi_a$ is equal to

$$F_\psi = \frac{SS_\psi}{MS_{\text{error}}} = \frac{SS_{\text{between}}}{MS_{\text{error}}} = \frac{(A-1)MS_{\text{between}}}{MS_{\text{error}}} = (A-1)F_{\text{omnibus}} \quad (2)$$

To have the Scheffé test contrast equivalent to the omnibus test, we need to reject the null hypothesis under the same conditions as the omnibus test. To reject the omnibus null hypothesis, $F_{\text{omnibus}}$ must be greater than or equal to $F_{\text{critical, omnibus}}$. Because $F$ is equal to $(A-1)F_{\text{omnibus}}$, then we reject the null hypothesis when

$$(A-1)F_{\text{omnibus}} > (A-1)F_{\text{critical, omnibus}} \quad (3)$$

and, therefore,

$$F_\psi > (A-1)F_{\text{critical, omnibus}}. \quad (4)$$

Consequently, the critical value to test all possible contrasts is

$$F_{\text{critical, Scheffé}} = (A-1)F_{\text{critical, omnibus}} \quad (5)$$

with

$$\nu_1 = A-1 \quad \text{and} \quad \nu_2 = A(S-1) \quad (6)$$

degrees of freedom. For our example, $F_{\text{critical, Scheffé}}$ is equal to

$$F_{\text{critical, Scheffé}} = (A-1)F_{\text{critical, omnibus}} = (4-1)\times 3.24 = 9.72 \quad (7)$$

with $\nu_1 = A - 1 = 4 - 1 = 3$ and $\nu_2 = A(S - 1) = 4(5 - 1) = 16$. An alternative approach is to correct the value of $F$ by dividing it by $(A - 1)$ and evaluating its probability according to a Fisher distribution with $(A - 1)$ and $A(S - 1)$ degrees of freedom (i.e., the number of degrees of freedom of the omnibus test).

The results of the Scheffé test for the contrasts in Table 3 are shown in Table 4. The third contrast, $\psi_3$, context before versus all other contexts, is the only significant contrast ($F_{\psi_3} = 19.92$, $p < .01$). This shows that memorization is facilitated only when the context information is presented before learning.

**Table 3** Post Hoc or A Posteriori Contrasts for the Bransford and Johnson (1972) Experiment

| | *Context Before* | *Partial Context* | *Context After* | *No Context* |
|---|---|---|---|---|
| $\psi_1$ | 1 | 1 | 1 | −3 |
| $\psi_2$ | 0 | 0 | 1 | −1 |
| $\psi_3$ | 3 | −1 | −1 | −1 |
| $\psi_4$ | 1 | −1 | 0 | 0 |

*Source:* Adapted from Bransford and Johnson (1972).

**Table 4** The Results of the Scheffé Test for the Contrasts in Table 3

| | $SS_\psi$ | $F_\psi$ | $p(F_{\text{Scheffé}})$ | *Decision* |
|---|---|---|---|---|
| $\psi_1$ | 12.15 | 5.17 | .2016 | ns |
| $\psi_2$ | 0.10 | 0.04 | F < 1 | ns |
| $\psi_3$ | 46.82 | 19.92 | .0040 | reject H0 |
| $\psi_4$ | 19.60 | 8.34 | .0748 | ns |

## *Pairwise Comparisons*

### Honestly Significant Difference Test

The honestly significant difference (HSD) test is a conservative test that uses Student's $q$ statistics and the Studentized range distribution. The range is the number of groups falling between groups (including the groups under consideration). For example, in the Bransford and Johnson experiment, the range of means between the largest and the smallest means is equal to 4 because there are four means going from 7.0 to 3.0 (i.e., 7.0, 4.2, 3.2, and 3.0).

To begin the HSD test, all pairwise differences are computed. These differences are shown in Table 6 for the Bransford and Johnson example. If the difference between two means is greater than the honestly significant difference, then the two conditions are significantly different at the chosen alpha level. The HSD is computed as

$$|M_{a+} - M_{a'+}| > \text{HSD} = q_{A,\alpha}\sqrt{\frac{1}{2}MS_{\text{error}}\left(\frac{1}{S_a} + \frac{1}{S_{a'}}\right)} \quad (8)$$

with a range equal to $A$ and $\nu = N - A$ degrees of freedom. For the Bransford and Johnson example, the HSD values are

$$\text{HSD}_{A,\alpha} = \text{HSD}_{4,\alpha=.05} = 2.37$$

for $p = .05$, and

$$\text{HSD}_{A,\alpha} = \text{HSD}_{4,\alpha=.01} = 3.59$$

for $p = .01$ with a range equal to $A = 4$ and $\nu = N - A = 20 - 4 = 16$ degrees of freedom. Using HSD, there are significant pairwise differences between the "context before" condition and the "partial context," "context after," and "no context" conditions.

### Newman–Keuls Test

The Newman–Keuls test is a sequential test in which $q_{\text{critical}}$ depends on the range of each pair of means. The Newman–Keuls test is the most popular a posteriori pairwise comparison test.

The Newman–Keuls test starts by computing the value of Student's $q_{\text{observed}}$ for the largest pairwise difference:

$$q_{\text{observed}} = \frac{M_{a+} - M_{a'+}}{\sqrt{MS_{\text{error}}\left(\frac{1}{S}\right)}} \quad (9)$$

(where $a$ and $a'$ are respectively the smallest and the largest means). This value is then evaluated with a Studentized distribution with a range of $A$ and with $\nu = N - K$.

In the Bransford and Johnson example, for the greatest pairwise difference, the $q_{\text{observed}}$ is equal to (cf. Equation 9)

$$q_{\text{observed}} = \frac{M_{1+} - M_{4+}}{\sqrt{MS_{\text{error}}\left(\frac{1}{S}\right)}} = \frac{4.0}{\sqrt{\frac{2.35}{5}}} = 5.83 \quad (10)$$

This value of $q_{\text{observed}} = 5.83$ needs to be compared to a value of $q_{\text{critical}}$ for a range of 4 and for $\nu = N - A = 20 - 4 = 16$ degrees of freedom. From the table of critical values for the Studentized range, we find that $q_{\text{critical}} = 4.05$ for $p = .05$ and 5.19 for $p = .01$; because $q_{\text{observed}} = 5.83$ is larger than $q_{\text{critical}} = 5.59$, we conclude that the "context before" and the "no context" conditions are significantly different at $p < .01$.

If the null hypothesis cannot be rejected for the largest difference, the test stops here. If the null hypothesis is rejected for the largest difference, the

**Table 5** Honestly Significant Difference Test

| | *Context Before* | *Partial Context* | *Context After* | *No Context* |
|---|---|---|---|---|
| | $M_{1+} = 7.0$ | $M_{2+} = 4.2$ | $M_{3+} = 3.2$ | $M_{4+} = 3.0$ |
| $M_{1+} = 7.0$ | | 2.8** | 3.8** | 4.0** |
| $M_{2+} = 4.2$ | | | 1.0 | 1.2 |
| $M_{3+} = 3.2$ | | | | 0.2 |

*Notes:* Pairwise differences between means for the Bransford and Johnson example. Differences greater than 2.37 are significant at the $p = .05$ level and are indicated by *. Differences greater than 3.59 are significant at the $p = .01$ level and are indicated by **.

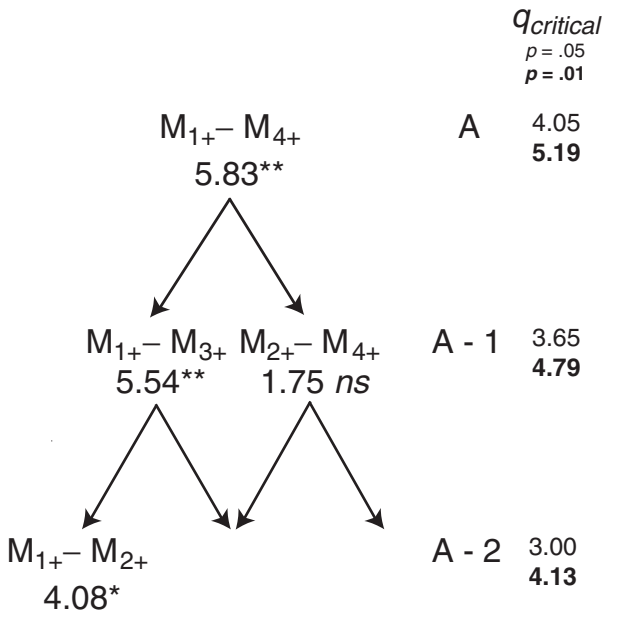



**Figure 2** Structure of Implication of the Pairwise Comparisons When $A = 4$ for the Newman-Keuls Test

*Notes:* Means are numbered from 1 (the largest) to 4 (the smallest). The pairwise comparisons implied by another one are obtained by following the arrows. When the null hypothesis cannot be rejected for one pairwise comparison, then all of the comparisons included in it are omitted from testing.

two differences with a range of $A - 1$ are examined. If the differences are nonsignificant, all other differences contained in that difference are not tested. If the differences are significant, then the procedure is reiterated until all means have been tested or have been declared nonsignificant by implication. This is shown for the Bransford and Johnson example by following the arrows in Figure 2. The results of the Newman–Keuls test are shown in Table 6. Again, we see significant differences between the "context before" condition and all other conditions.

### Duncan Test

The Duncan test follows the same general sequential pattern as the Newman–Keuls test. The difference is that the critical values for the test come from a different table. The only difference between the two tests is that the Duncan test uses the Fisher $F$ distribution with a Šidák correction. The Šidák correction is computed using the Šidák inequality

$$\alpha[PC] = 1 - (1 - \alpha[PF])^{\frac{1}{\text{Range}-1}} \qquad (11)$$

where $\alpha[PC]$ is the $\alpha$-level per comparison and $\alpha[PF]$ is the $\alpha$-level per family of comparisons. For the Bransford and Johnson example with $A = 4$ conditions,

**Table 6** Newman-Keuls Test

| | *Context Before* | *Partial Context* | *Context After* | *No Context* |
|---|---|---|---|---|
| | $M_{1+} = 7.0$ | $M_{2+} = 4.2$ | $M_{3+} = 3.2$ | $M_{4+} = 3.0$ |
| $M_{1+} = 7.0$ | | $2.8^{*}$ | $3.8^{**}$ | $4.0^{**}$ |
| $M_{2+} = 4.2$ | | | 1.0 | 1.2 |
| $M_{3+} = 3.2$ | | | | 0.2 |

*Notes:* Pairwise differences between means for the Bransford and Johnson example. Differences significant at the $p = .05$ level and are indicated by $^{*}$. Differences significant at the $p = .01$ level and are indicated by $^{**}$.

**Table 7** Duncan Test

| | *Context Before* | *Partial Context* | *Context After* | *No Context* |
|---|---|---|---|---|
| | $M_{1+} = 7.0$ | $M_{2+} = 4.2$ | $M_{3+} = 3.2$ | $M_{4+} = 3.0$ |
| $M_{1+} = 7.0$ | | $8.3403^{*}$ | $15.3619^{**}$ | $17.0213^{**}$ |
| $M_{2+} = 4.2$ | | | 1.0638 | 1.5320 |
| $M_{3+} = 3.2$ | | | | 0.0425 |

*Notes:* $F$ values for the Bransford and Johnson example. Significant $F$ values at the $\alpha[PF] = .05$ level are indicated by *, and at the $\alpha[PF] = .01$ level by $^{**}$.

$$\begin{aligned}\alpha[PC] &= 1 - (1 - \alpha[PF])^{\frac{1}{\text{Range}-1}} \\ &= 1 - (1 - .05)^{\frac{1}{3}} \\ &= 0.0170.\end{aligned} \tag{12}$$

The $q$ values computed for the Newman–Keuls test can be changed into $F$ values using the following fomula:

$$F_{\text{range}} = \frac{q^2}{2}. \tag{13}$$

For the Bransford and Johnson example, this results in the $F$ values shown in Table 7.

For the first step, the critical $F$ values are 7.10 for an $\alpha[PF]$ of .05 and 12.45 for an $\alpha[PF]$ of .01 with range $= A = 4$ and $\nu_2 = N - A = 16$ degrees of freedom. Then, the same recursive procedure as the Newman–Keuls test is followed and results in the pattern shown in Table 7.

*Lynne J. Williams and Hervé Abdi*

***See also*** Analysis of Variance (ANOVA); Bonferroni Procedure; Contrast Analysis; Duncan's Multiple Range Test; Fisher's Least Significant Difference Test; Multiple Comparison Tests; Newman–Keuls Test and Tukey Test; Pairwise Comparisons; Post Hoc Comparisons; Scheffé Test; Tukey's Honestly Significant Difference

**Further Readings**

Abdi, H., Edelman, B., Valentin, D., & Dowling, W. J. (2009). *Experimental design and analysis for psychology*. Oxford, UK: Oxford University Press.

Bransford, J. D., & Johnson, M. K. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior, 11*, 717–726.

Hochberg, Y., & Tamhane, A. C. (1987). *Multiple comparison procedures*. New York: Wiley.

Jaccard, J., Becker, M. A., & Wood, G. (1984). Pairwise multiple comparison procedures: A review. *Psychological Bulletin, 94*, 589–596.

Miller, R. G. (1981). *Simultaneous statistical inferences*. Berlin: Springer Verlag.

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach*. Cambridge, UK: Cambridge University Press.

# Power

The power of a statistical test is the probability that the selected test will appropriately reject the null hypothesis (i.e., when an alternative hypothesis is true). That is, it refers to the likelihood that the test will not make a Type II error (false negative rate or $\beta$). Because power is equal to $1 - \beta$, as Type II error decreases, power increases. Statistical power is influenced by statistical significance, effect size, and sample size. All of these factors are taken into consideration when completing a power analysis.

## Primary Factors That Influence Power

Several factors influence power, or the ability to detect significant results if they exist: statistical significance, effect size, and sample size. Each of these terms is described in this section.

### *Statistical Significance*

The significance level ($\alpha$) is the probability of rejecting a null hypothesis that is true. The power of the test ($1 - \beta$) is the probability of correctly rejecting a false null hypothesis. Therefore, as significance levels increase, so does power. One way to increase power is to use a larger significant criterion, which increases the chance of obtaining a statistically significant result (rejecting the null hypothesis). However, doing so increases the risk of obtaining a statistically significant result when the null hypothesis is true (i.e., false positive, or Type I error). A commonly used, albeit arbitrary, significance level is $\alpha = .05$ which signifies there is a 5% probability that a researcher will incorrectly detect a significant effect when one does not actually exist.

### *Effect Size*

Effect size refers to the magnitude of the effect of interest in the population. Larger effects are easier to detect than small effects. Thus, power to detect a significant effect increases as the magnitude of the effect increases. For example, a researcher may be likely to detect a very large difference between two groups but may have

more difficulty detecting a small difference. In the latter case involving very small effects, the probability of a Type II error, which refers to not finding a significant difference when one actually exists (i.e., a false negative, or $\beta$), is high. For instance, if the likelihood of a false negative ($\beta$) is .80, and power is equal to $1 - \beta$, then power in this case is .20. So, in general, the power to detect small effects is low, although it can be increased by manipulating other parameters, such as sample size.

### *Sample Size*

Sample size refers to the number of observations ($n$). In the case of human sciences, this often means the number of people involved in the study. In power analysis, the most frequently asked question is how many observations need to be collected in order to achieve sufficient statistical power. When sample size is large, variation within the sample (standard error) becomes smaller and makes standardized effect size larger. Note that there may be times when recommended sample size (resulting from power analysis) will be inadequate (see example below). Although increasing sample size may increase power, there is recognition that too many observations may lead to mistaken detection of trivial effects that are not clinically significant. By contrast, if too few observations are used, a hypothesis test will be weak and less convincing. Accordingly, there may be little chance to detect a meaningful effect even when it exists (Type II error).

## Power Analysis

Power analysis is typically required by funding agencies and ethics boards because it helps determine number of observations or participants necessary for recruitment. Power analysis can be done either before (a priori or prospective) or after (post hoc or retrospective) data are collected. A priori power analysis is typically used to determine an appropriate sample size to achieve adequate power and requires determination of the test model that will be used for data analysis (e.g., $t$ test). Post hoc power analysis is used to determine the magnitude of effect size of the observed sample compared to the population. Generally, a priori power analysis is preferred and done during the design stage of the study.

Although there are no formal standards for power, power of .80 is usually the accepted standard for hypothesis testing.

## Calculating Power

Because power is a function of $\alpha$, $n$, and effect size, these should be predetermined prior to computing power. Most likely, power calculations are computed in order to predetermine $n$. For example, a researcher may conduct a power analysis to determine how many people will be needed to detect a significant ($\alpha = .05$) effect with power equal to .80 for a specified statistical test (e.g., analysis of variance). The goal is to achieve a balance of the components that allows the maximum level of power to detect an effect if one exists. In most cases, power is obtained by simple computations, or by using power tables.

The following are additional resources for power calculations:

*nQuery Advisor Release 4.0*—This software is used for sample size estimate and power calculations and contains extensive table entries and many other convenient features.

*SamplePower(r) 1.2*—SamplePower is available from SPSS, an IBM company, and arrives at sample size for a variety of common data analysis situations.

*G*Power*—G*Power allows you to calculate a sample size for a given effect size, alpha level, and power value.

*Hoa T. Vo and Lisa M. James*

***See also*** Effect Size, Measures of; Experimental Design; Sample Size; Significance Level, Concept of; Significance Level, Interpretation and Construction; Type I Error; Type II Error

### Further Readings

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159.
Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole.
Thomas, L., & Krebs, C. J. (1997). A review of statistical power analysis software. *Bulletin of the Ecological Society of America, 78*(2), 126–139. Retrieved from http://www.zoology.ubc.ca/~krebs/power.html

**Websites**

*G*Power:* http://www.psycho.uni-duesseldorf.de/aap/projects/gpower
SPSS, *SamplePower(r) 1.2:* http://www.spss.com/spower/research.htm
Statistical Solutions, *nQuery Advisor Release 4.0:* http://www.statsol.ie

# Power Analysis

A primary objective of many studies is to demonstrate a difference between two or more treatments under investigation. Power analysis, also referred to as sample size calculation, plays an important role in ensuring that a sufficient number of subjects are enrolled for answering the question of interest. Specifically, it is important to design a study in such a way that it will have a high probability of showing a difference when a difference truly exists and a low probability of showing a difference when none exists. If the sample size is too small, the study will be underpowered and may lead to discarding a potentially useful treatment. Such underpowered studies often lead to great confusion in the literature because they are often perceived as negative studies, whereas in actuality, this is not the case. Furthermore, from an investigator standpoint, cost and effort are devoted to a study that fails to prove or disprove the question of interest. Although studies with larger sample sizes than required are not affected by these same concerns, such studies are wasteful of important study resources that might have been directed elsewhere. Correspondingly, sample size calculation should play an important role during the planning stage of any study.

## Basic Principles for Sample Size Calculation

Sample size calculation is usually performed based on some statistical criteria controlling the Type I and Type II errors (see Table 1).

### *Type I Error*

The *Type I error* ($\alpha$) is the probability of rejecting the null hypothesis when it is true:

$$\begin{aligned}\alpha &= \Pr\{\text{Type I Error}\} \\ &= \Pr\{\text{Reject } H_0 \text{ when true}\}\end{aligned}. \quad (1)$$

For example, suppose that there are two groups of observations, where $x_i$ and $y_i$ ($i = 1, \ldots, n$) correspond to subjects receiving treatment and control, respectively. Assume that $x_i$ and $y_i$ are *independent* and *normally distributed* with means $\mu_1$ and $\mu_2$, respectively, and a common variance of $\sigma^2$. Let $\delta = \mu_1 - \mu_2$ represent the difference between the two means for the two groups. A test of *equality* attempts to show that one treatment is more effective than another:

$$H_0 : \delta = 0 \text{ vs. } H_a : \delta \neq 0 \ . \quad (2)$$

In this, a Type I error refers to the probability of incorrectly concluding that the population means differ when there is actually no difference. The most common approach is to specify $\alpha = 0.05$ Note that this implies that we would expect to reject the null hypothesis approximately 5% of the time when it is not true (there is no effect). However, this is often chosen more by convention than design. Any level of Type I error can be selected for any given study.

**Table 1** Possible Outcomes for a Trial

| | *Truth* | |
|---|---|---|
| *Conclusion* | *No Treatment Benefit* | *Treatment Benefit* |
| *Evidence of Treatment Effect* | Type I Error (False Positive) | Correct Result (True Positive) |
| *No Evidence of Treatment Effect* | Correct Result (True Negative) | Type II Error (False Negative) |

### *Type II Error*

The *Type II error* ($\beta$) is the probability of not rejecting the null hypothesis when it is false:

$$\begin{aligned}\beta &= \Pr\{\text{Type II Error}\} \\ &= \Pr\{\text{Fail to reject } H_0 \text{ when false}\}\end{aligned}. \quad (3)$$

In the previous example, this refers to the probability of not concluding that the population means differ when there is actually a difference. The *power* of a test is defined as the probability of rejecting the null hypothesis given some assumed effect. Hence, when a true effect is assumed, there is a clear inverse relationship between the power and the probability of a type II error:

$$\begin{aligned}\text{Power} &= 1 - \beta \\ &= \Pr\{\text{Reject } H_0 \text{ when false} \,|\, \text{assumed effect}\}\end{aligned} \quad (4)$$

Ideally, one would like to make the probability of both types of errors as small as possible. However, with a fixed sample size, the Type I error increases as the Type II error decreases and vice versa. The only approach that allows simultaneously decreasing both types of errors is to increase the sample size.

### *Additional Items*

In addition to the error rates described above, there are several additional items that must be specified in order to perform such calculations. Each is addressed below.

#### Type of Data

The sample size fomula must take into account the type of data that will be observed in any given trial. A continuous response consists of an outcome, such as blood pressure, that can be measured on a continuous scale. A dichotomous response consists of an outcome that can take one of two conditions for each subject in the trial (success or failure, presence or absence). Time-to-event data occur when not all subjects will have an event during the time frame of the study. Subjects who do not have the event are censored for the purposes of analysis. Although there are other types of variables, sample size estimates for response variables that do not fall into these categories usually can be approximated by techniques for one of them.

#### Type of Test

Different types of tests require different sample sizes to achieve a desired power. Hence, it is vitally important to ensure that the sample size calculation parallels the planned primary analysis. Hence, the hypothesis of interest must be clearly stated when performing a sample size calculation. An example of a test for equality was described in Equation 2. A test of *superiority* attempts to show that a test drug is more effective than an active agent or standard therapy:

$$H_0 : \delta \leq \varepsilon \text{ vs. } H_a : \delta > \varepsilon, \quad (5)$$

where $\varepsilon$ is the superiority margin. A test of *noninferiority* attempts to show that a test drug is as effective as an active agent or standard therapy:

$$H_0 : \delta \leq -\varepsilon \text{ vs. } H_a : \delta > -\varepsilon, \quad (6)$$

where $\varepsilon$ is the noninferiority margin. A test of *equivalence* attempts to show no difference of clinical importance between two treatments:

$$H_0 : |\delta| \geq \varepsilon \text{ vs. } H_a : |\delta| < \varepsilon\ , \quad (7)$$

where $\varepsilon$ is the equivalence margin. In the remainder of this entry, the focus primarily is on tests of equality.

#### Precision

A more precise method of measurement (i.e., smaller variance) will permit detection of any given effect with a smaller sample size. A study with a small sample size will have more uncertainty and will show statistically significant differences only if there is a large difference between the two groups.

#### Nuisance Parameters

Nuisance parameters are those parameters that must be specified in order to conduct the calculation, but are not associated with the primary effect of interest. Some common examples include the variance with continuous data, the control group

event rate with binary or time-to-event data, and the average follow-up time for time-to-event data.

Given these values, there are usually three other parameters of interest: the treatment effect, the power, and the sample size. If we are given all but one of these parameters, the other parameter is completely determined. Correspondingly, there are three general types of power analyses.

A *sample size justification* is conducted to provide evidence that a desired effect of interest can be detected for some fixed level of power and sample size. Such calculations are often used when the sample size has been restricted due to budget and/or time constraints. In general, this should be used only as a last resort because it becomes difficult to determine whether the detectable effect is clinically or scientifically meaningful.

A *power analysis* is conducted to determine the power of a study with a fixed sample size and detectable effect of interest. For such analyses, the choice of a treatment effect of interest is critical. If the effect of interest is small, a large sample size will be required to have a high probability of detecting the effect at a specified level of power. It is also critically important to ensure that specified treatment effects have both clinical and statistical meaning. For instance, it is *possible* to design a study to detect a reduction of systolic blood pressure by 0.1 mmHg. However, showing such a small difference in a study would likely not be of clinical importance and would not change practice.

A *sample size calculation* is conducted to determine the sample size required to achieve some prespecified target power, given a detectable effect of interest. The desired level of power is typically set at 80% or 90% for planning purposes. Power curves (see Figure 1) are often useful because the choice of sample size often involves a trade-off between the desire to obtain greater power versus the added cost of the study. Power curves typically have a sigmoidal shape, increasing with both sample size and the effect of interest. In general, because trials designed with 80% power lie closer to the shoulder of the curve, they are more susceptible to inaccuracies in the design parameters than trials designed with 90% power.

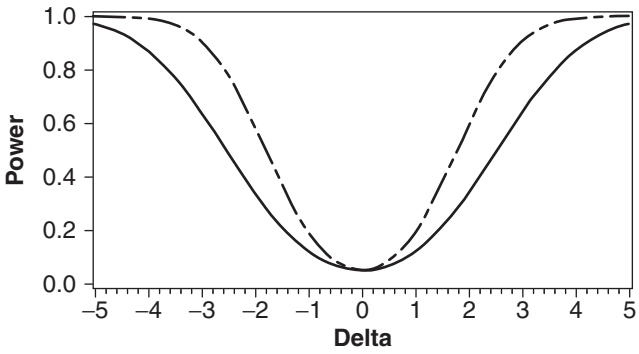



**Figure 1** An Example of a Power Curve

Note that we cannot separate power from either the size of the study or the magnitude of the effect. Hence, the following statement is ambiguous: "The trial has 90% power." Rather, all three values must be discussed simultaneously: "With 500 subjects per group, the trial has 90% power to detect a decrease of 10 mmHg in blood pressure due to the new treatment at the 5% significance level." A good rule of thumb is that sufficient explanation and documentation should be provided so that an independent party can duplicate the calculation

## Sample Size Calculations for Two Group Comparisons

### *Continuous Response*

Consider the hypothesis described in Equation 2. When $\sigma^2$ is known, we reject $H_0$ at the $\alpha$ level of significance if

$$Z = \left| \frac{\sqrt{n/2}(\overline{X} - \overline{Y})}{\sigma} \right| > z_{\alpha/2}, \tag{8}$$

where

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x \ , \ \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \tag{9}$$

and $z_p$ is the $p$th quantile of the standard normal distribution (see Figure 2). Suppose that we are interested in determining the sample size required to achieve power $1 - \beta$ for this test under the alternative hypothesis that $\mu_1 = \mu_2 + \delta$ (where $\delta$ is a positive clinically meaningful treatment difference). Under such an alternative, the distribution

of the test statistic is shifted away from zero (see Figure 2 for an example with $\delta = 1$). The power is then the area under the alternative distribution that lies in the rejection region, which is approximately

$$\Pr\left\{Z > z_{\alpha/2} - \frac{\delta\sqrt{n/2}}{\sigma}\right\} = \Phi\left(\frac{\delta\sqrt{n/2}}{\sigma} - z_{\alpha/2}\right). \quad (10)$$

As a result, the sample size per group can be obtained by solving the following equation:

$$\frac{\delta\sqrt{n}/2}{\sigma} - z_{\alpha/2} = z_{\beta}$$
$$n = \frac{2\sigma^2\left(z_{\alpha/2} + z_{\beta}\right)^2}{\delta^2}. \quad (12)$$

As an example, consider a study of a new treatment to reduce blood pressure. Suppose that a 10 mmHg difference would be clinically significant ($\delta = 5$) and the standard deviation is estimated to be 25 mmHg ($\sigma = 25$). Suppose we want to compute the required sample size to detect this effect based on a two-sided test with equal sample sizes, a Type I error rate of 5% ($\alpha = 0.05$), and a target level of 90% power ($\beta = 0.10$). Substituting these values into the fomula gives

$$n = \frac{2(1.96 + 1.28)^2(25)^2}{(5)^2} = 524.9. \quad (12)$$

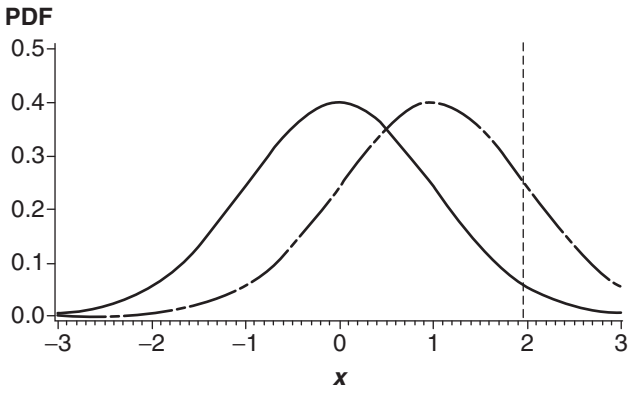



**Figure 2** Normal Distribution Under the Null ($\mu = 0$) and Alternative ($\mu = 1$) Hypothesis

Rounding up suggests a required sample size of approximately 525 per group. If the estimated standard deviation were actually 30, the study would require a sample size of approximately 760 per group. This reinforces the importance of being somewhat conservative when determining the required sample size.

When $\sigma^2$ is not known, it can be replaced by $s^2$ (the sample variance), which results in the usual two-sample $t$ test. When $n$ is sufficiently large, the normal distribution can be used to approximate the sample size

$$n = \frac{s^2\left(z_{\alpha/2} + z_{\beta}\right)^2}{\delta^2}. \quad (13)$$

Otherwise, one should use a fomula based on a $t$-distribution.

### *Dichotomous Response*

Suppose that we are still interested in comparing a treatment to a control, but the outcome of interest is whether or not an event has occurred. Let $p_T$ represent the proportion of failures on treatment, $p_C$ the proportion of failures on control, and $\delta = p_T - p_C$. The test of interest can be conducted by stating the hypothesis as in Equation 2. The total sample size required in each group is, approximately,

$$n = \frac{\left\{z_{\alpha/2}\sqrt{2\overline{p}(1-\overline{p})} + z_{\beta}\sqrt{p_T(1-p_T) + p_C(1-p_C)}\right\}^2}{(p_T - p_C)^2}, \quad (14)$$

where

$$\overline{p} = (p_T + p_C)/2. \quad (15)$$

In general, the variance is largest when $p = 0.5$ and smallest when $p$ is near zero or one. Thus, if one has no idea about the true value of $p_C$ (a nuisance parameter), then one can conservatively use $p_C = 0.5$ in the variance fomula for the sample size calculations.

*Time-to-Event Response*

With time-to-event data, we need to have a sufficient number of *events* rather than total observations. Hence, sample size calculations for comparing two survival curves consist of a two-step process. First, the required number of events must be calculated. Assuming equal subjects in the two groups, the required number of events to detect a specified log hazard ratio ($\beta_*$) for a given study is given by

$$\text{required \# of events} = \frac{4\left(z_{\alpha/2} + z_{\beta}\right)^2}{\beta_*^2}. \qquad (16)$$

Second, a required number of patients must be computed that is expected to provide the required number of events during the time frame of the study. This calculation requires estimating the probability of the event occurring over the course of the study. Clearly, the longer the accrual and follow-up time, the longer a patient will be in the study and the greater the probability (on average) that an event will occur. Once the probability of the event has been determined, the required number of subjects can be found from

$$\text{required sample size} = \frac{\text{required \# of events}}{\Pr(\text{event})}. \qquad (17)$$

## Sample Size Calculations for More General Settings

Although the fomulas described above for two group comparisons adequately cover a large number of studies, there are obviously many studies involving more than two groups. Although such settings add to the complexity, the underlying basic principles for sample size calculation described above remain valid. Sample size calculations for the general linear univariate model setting are described in detail in Chapter 17 of Keith E. Muller and Bethel A. Fetterman's *Regression and ANOVA: An Integrated Approach Using SAS Software*. Muller and Paul W. Stewart describe similar results for the general linear multivariate and linear mixed-model setting. Ralph G. O'Brien describes a general procedure based on exemplary data for computing the power of a likelihood ratio test in a variety of settings. Many of these features have been incorporated into the POWER and GLMPOWER procedures recently released by SAS Institute.

## Software

There are an increasing number of software options for sample size calculations. These include free software packages, such as Russell Lenth's Java applets for power and sample size website, PS: Power and Sample Size Calculation, and POWERLIB, as well as commercial packages, such as PASS (NCSS), NQuery Advisor (Statistical Solutions), and PROC POWER and PROC GLMPOWER (SAS Institute).

In general, the free packages tend to handle a small number of situations very well but do not have the graphical interfaces available with commercial software.

Despite the range of statistical programs available, it is not uncommon to encounter a situation that is not found in any available programs. Thus, it may be necessary to write simulation code to assess the characteristics of the study design of interest. In such settings, it is very important to consider what range of scenarios should be simulated. As with standard power calculations, one must provide sufficient information that would allow one to reproduce the simulations if necessary. However, because it is extremely difficult for a reviewer to completely reproduce such simulations, it is imperative that proper documentation be included in any power analysis that is conducted via simulation.

## Other Issues

Many fomulas for sample size calculation are based on large sample approximations. With large, randomized clinical trials, this assumption is typically valid. For many individuals, examples covered by large sample approximations are the only ones covered in an introductory statistics course. This gives the false impression that sample size calculations are always quick and simple. In general, this is not the case. Situations that require simulations were described above. For smaller studies,

such as those conducted in Phase I or II, large sample approximations may not be appropriate and more exact power calculations may be required.

Sample size calculation provides the number of *evaluable* subjects that are required in order to achieve some desired level of power. If dropouts are expected, the number of subjects enrolled may need to be larger in order to ensure a sufficient number of evaluable subjects.

Small changes in the nuisance parameters, which must be specified for a sample size calculation, may yield large changes in the required sample size. Internal pilot designs can be used to reassess nuisance parameters after the study has started and adjust the sample size up or down to preserve the desired level of target power.

Although not described in this entry, there are alternative methods for choosing a sample size. A *precision analysis* chooses a sample size such that there is a desired precision at a fixed confidence level (i.e., fixed Type I error). Recent research has suggested that the optimal approach to choosing a sample size is to account for both precision and power when designing a study. However, despite this recommendation, this approach does not seem very common in practice to date.

### Conclusion

The size of a study should be considered early in the planning phase. Before emoverlineking on a research plan, researchers should have sufficient statistical power to detect differences between groups considered to be of clinical interest. Therefore, calculation of an appropriate sample size and power is an essential part of planning. In general, the effort devoted to a sample size calculation should reflect the importance of selecting an appropriate sample size for a given study. Hence, researchers should devote as much effort to study planning and selection of sample size as they plan to devote to the statistical analysis at the end of the study.

*Christopher S. Coffey*

***See also*** Sample Size; Sample Size Planning

#### Further Readings

Chow, S. C., Shao, J., & Wang, H. (2008). *Sample size calculations in clinical research*. Boca Raton, FL: Chapman & Hall/CRC.

Jiroutek, M. R., Muller, K. E., Kupper, L. L., & Stewart, P. W. (2003). A new method for choosing sample size for confidence interval-based inferences. *Biometrics, 59,* 580–590.

Muller, K. E., & Fetterman, B. A. (2002). *Regression and ANOVA: An integrated approach using SAS software*. Cary, NC: SAS Institute.

Muller, K. E., & Stewart, P. W. (2006). *Linear model theory: Univariate, multivariate, and mixed models*. Hoboken, NJ: Wiley.

O'Brien, R. G. (1998). A tour of UnifyPow: A SAS module/macro for sample size analysis. *Proceedings of the 23rd SAS Users Group International Conference* (pp. 1346–1355). Cary, NC: SAS Institute.

Wittes, J., & Brittain, E. (1990). The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine, 9,* 65–72.

#### Websites

Lenth, R. Java applets for power and sample size: http://www.stat.uiowa.edu/~rlenth/Power/index.html

POWERLIB: http://ehpr.ufl.edu/muller

PS: Power and Sample Size Calculation, Vanderbilt University: http://biostat.mc.vanderbilt.edu/wiki/bin/view/Main/PowerSampleSize

## Pragmatic Study

A pragmatic study focuses on an individual decision maker within an actual real-world situation. The process of undertaking a pragmatic study is first to identify a problem and view it within its broadest context. This leads to research inquiry, which seeks to better understand and ultimately solve the problem. Finally, the research findings often result in policy suggestions, new environmental initiatives, or social change.

### Real-World Research

Is it important to debate the philosophy behind our research and the theoretical basis of our research approach? A pragmatic study would be only tangentially interested in such deeper questions, as it puts practical solutions above

philosophical discussions. Pragmatic studies often draw upon mixed-methods approaches. Both qualitative and quantitative methods could be employed—whatever methods provide a relevant approach to a given research question. This enables a researcher to develop a holistic analysis to fully incorporate numerous relevant factors into the study. Pragmatic studies are inductive, moving from a complex problem to a general theory of understanding in order to improve a given situation.

Pragmatism is based on understanding human experience. So, pragmatic studies often seek to understand the multiple factors involved in people's actions in a given situation. Thus, pragmatists acknowledge that their inquiry will not lead to certainty because, in theory, nothing in the world is certain. William James, a central figure in this philosophy, noted that a pragmatic study avoids abstract, fixed principles and does not pretend there is only one final truth. Instead, a pragmatic study defines terms by their application to human experience. Most pragmatic researchers are motivated to conduct research in order to address problems in the real world.

In undertaking pragmatic studies, a researcher believes that the research itself should be used to solve problems and improve human and ecological conditions. Many pragmatists would agree with Gilbert F. White's research goals, which are driven by solving problems that truly affect people and seeking to translate research results into action. This concept exemplifies the basic tenet of a pragmatic study: to conduct research whose results can be translated into practical ends. This often involves policy recommendations or other real-world solutions.

## Research Goals

There are four main goals that are often employed to guide pragmatic research: Accept chaos in interrelationships among variables; seek an understanding based on human experience; view a problem as a complex *problematic situation;* and promote activism, democracy, and policy fomulation.

First, any given setting provides uncertainty that people define and attempt to resolve; thus, research variables are interdependent. A pragmatic study accepts that any setting is precarious and, indeed, assumes it is a very real feature of every situation, according to pragmatist John Dewey. So, for example, in order to address current ecological problems, which are often an outcome of our attempts to control nature, pragmatism accepts the premise that nature is basically unpredictable. Just as physical scientists question the suitability of linear prediction fomulas for weather forecasting, social scientists theorize that interactions between society and the environment assume a nonlinear, chaotic relationship. Pragmatic studies allow for such views by accepting a systems approach to our individual and societal relationships with nature. This provides research flexibility and the ability to select the best possible methods, because pragmatism does not require a linear investigation that seeks one resultant "truth."

A second goal of the pragmatic study is to embrace human experience as the basis for decision making. According to Dewey, experience is what we do, how we learn, and what happens to us. Pragmatist Charles Peirce noted that we should look at general long-term effects and experiences, not just short-term specific ones. Furthermore, knowledge is valued for its usefulness, and understanding of a situation must be gained through one's own experiences or inferred from other people's experiences. Pragmatism focuses on tangible situations, not abstractions. By investigating human behavior, we actually attempt to understand human experiences. Thus, pragmatic inquiry is only instrumental; it is not an end in itself. The research aim is not to seek an absolute truth, because none exists; rather, it is to fomulate policies and aid in improving society. Abstract questions are rarely posed because they are not often directly relevant to a specific research case. So, pragmatic studies investigate human experience to understand the "truth" of what works best in a given situation.

Third, pragmatists seek to improve *problematic situations* rather than test narrow research hypotheses removed from their contexts. Pragmatic studies view a problem not as an isolated event, but rather in its full context. A *problematic situation* is investigated through "decision elements" (e.g., social, technological, ecological, economic, perception, and spatial factors) that can be explored concurrently and interactively. There is no fixed progression, and new elements may be discovered

to better define the problem. This again points to the use of mixed-methods approaches in pragmatic studies. Qualitative techniques are complementary to quantitative methods and are effective in searching out the complexity of human behavior. A pragmatic researcher seeks to transform a problem by investigating its complex interrelated elements in order to better understand the entire situation. The goal is to present alternatives and to take appropriate action. This leads to social activism and appropriate policies.

Many problems, and thus research questions, are firmly embedded in our society, so the pragmatic aim of transforming a *problematic situation* often confronts the status quo. Pragmatic approaches acknowledge that scholars may need to confront entrenched social and political constraints. This leads to the fourth goal: to contribute a unique perspective that links individual decision making to societal change through public participation. Academic inquiry can inform public choice if research is practical and applied to issues in the real world. A pragmatic study can provide a link to other philosophies through an emphasis on individual decision makers at a grassroots level acting to promote political and social change. A pragmatic study would not stress class, gender, or race differences, but rather seek to encourage each individual to be informed and take part in a truly democratic process. Pragmatists like Dewey believed that our democratic society requires diversity and a complex understanding of the world. Pragmatic behavioral research seeks to understand and reinforce the capacities of individual people for action. Individual and community empowerment can improve a *problematic situation* and eventually lead to social benefits.

## Applications

Pragmatism is concerned with understanding and resolving problems that occur in our uncertain world. Thus, environmental topics are particularly conducive to such applied goals, as natural resource management and environmental policy invite practical action. Historically, the work of Gilbert F. White and his students represent one school of thought that focuses on practical research on natural hazards and a broad range of natural resource themes. In addition, pragmatic studies are now also common in many areas of research, including health care and information technology, among others.

## Opportunities for Pragmatic Studies

A pragmatic study can address social and environmental perspectives that allow researchers to investigate human experience, human adjustment to various processes, and the subsequent range of choice that individuals and society identify. Pragmatic researchers often speak publicly to raise concerns, inform people, and encourage cooperation. In addition, this approach is particularly valuable at a grassroots level for encouraging ideas and movements regarded as outside the mainstream. For more researchers to employ behavioral pragmatism, it must be recognized as a valid philosophical approach for the social sciences. Most adherents to this approach prefer to spend their time on applied research projects that deal with human problems, rather than partake in debates on philosophical/theoretical concerns. This may explain why pragmatic studies are not well-known in the natural or social sciences. Yet pragmatism allows open and comprehensive investigation, as there are no theoretical constraints that limit the inquiries. Thus, researchers should be aware of and use a pragmatic approach to understanding complex, real-world problems.

*Leslie A. Duram*

***See also*** Action Research; Applied Research; Case Study; Field Study; Focus Group; Interviewing; Mixed Methods Design

### Further Readings

Dewey, J. (1958). *Experience and nature*. New York: Dover.

James, W. (1907). *Pragmatism: A new name for some old ways of thinking*. New York: Longman & Green.

Light, A., & Katz, E. (Eds.). (1996). *Environmental pragmatism*. London: Routledge.

Wescoat, J. L. (1992). Common themes in the work of Gilbert White and John Dewey: A pragmatic appraisal. *Annals of the Association of American Geographers, 82,* 587–607.

White, G. F. (1972). Geography and public policy. *Professional Geographer, 24,* 101–104.

# Precision

The term *precision* refers to how precisely an object of study is measured. Measurements of an object can be made with various degrees of precision. The amount of precision will vary with the research requirements. For example, the measurement of baby ages requires a more precise measurement than that of adult ages; baby ages are measured in months, whereas adult ages are measured in years.

The term *precision* also refers to the degree to which several measurements of the same object show the same or similar results. In this regard, precision is closely related to reliability. The closer the results of measurements, the more precise the object measurement is. Measurement with high precision is very likely to produce the same and predictive results.

This entry focuses on precision with regard to research requirements, accuracy, and reliability.

## Research Requirements and Precision

Measurements can be made with various degrees of precision, and how precisely a measurement should be made depends on the research requirement. Precise measurements are always better than imprecise ones, but more precise measurements are not always superior to less precise ones. The description of a residential location as "123 5th Street" is more precise than "inner-city neighborhood." The description of a historic house built "in March 1876" is much more precise than "in the late 19th century."

More precise measurements such as "123 5th Street" are not always necessary or desirable. If the description of residential locations such as inner city, inner suburbs, or suburbs satisfies the research requirement, the more precise measurements (i.e., complete residential addresses) necessitate that the researchers identify whether these addresses are located in the inner city, inner suburbs, or suburbs.

An understanding of the degree of precision is also required to direct an efficient data collection. For example, the work experience of people is measured in years or months, but not in hours. The preparation time of students for final exams is measured in hours, but not in minutes or seconds. Any attempt that measures the work experience of people in hours or the student preparation time for final exams in minutes is wasted.

Less precise measurements are not always inferior to more precise ones. In the case of people's work experience, the measurement in years or months is better than hours. Similarly, the measurement of student preparation time for final exams in hours is better than minutes.

## Precision and Accuracy

Precision is an important criterion of measurement quality and is often associated with accuracy. In experimental sciences. including social and behavioral sciences, there is low precision, low accuracy; low precision, high accuracy; high precision, low accuracy; and high precision, high accuracy. The measurement with high precision and high accuracy is certainly a perfect measurement that can hardly be made. The best measurement that can be made is to come as close as possible within the limitations of the measuring instruments.

It is easier to be accurate when the measurement does not aim at producing a precise result. In the meantime, if the measurement aims at obtaining a precise result, then it is likely to produce an inaccurate result. The most commonly used illustration to exemplify the difference between precision and accuracy is dart throwing. Dart throwing is neither precise nor accurate when the darts are not clustered together and are not near the center of the target. It is precise but inaccurate when the darts are clustered together but are not near the center of the target. It is accurate but not precise when the darts are not clustered together but their average positions are the center of the target. It is both precise and accurate when the darts are clustered together in the center of the target.

## Precision and Reliability

Precision, particularly in the physical sciences, is similar to reliability and is also called reproducibility and consistency. Precise measurements are repeatable and reliable measurements that have similar results. Several conditions are required to produce precise measurements, including proper use of reliable equipment and careful and

consistent measurement procedures. In addition, the use of the same instrument from the same manufacturer will likely produce precise measurements, as opposed to switching to another instrument from a different manufacturer.

Precision is also a function of error due to chance variability or random errors. Less precise measurements have more random errors than do more precise ones. Four possible sources of random errors in making measurements are operator variability, method variability, instrument variability, and subject variability. Operator variability refers to measurement variability due to different operators, such as word choices in an interview or skill in operating a measuring instrument. Method variability refers to the variability in measurement due to use of the wrong measuring procedure. Instrument variability refers to measurement variability due to use of an unreliable instrument, such as an instrument that has not been calibrated recently. Subject variability refers to measurement variability that is due to intrinsic biologic fluctuations, such as fluctuations in mood.

*Deden Rukmana*

***See also*** Accuracy in Parameter Estimation; Levels of Measurement; Reliability; Validity of Measurement

**Further Readings**

Babbie, E. (2007). *The practice of social research* (11th ed.). Belmont, CA: Thomson/Wadsworth.

Connelly, L. M. (2008). Accuracy and precision. *MEDSURG Nursing, 17*(2), 99–100.

Dillman, D. A. (2007). *Mail and Internet surveys: The tailored design method* (2nd ed.). Hoboken, NJ: Wiley.

Hulley, S. B., Martin, J. N., & Cummings, S. R. (2007). Planning the measurements: Precision and accuracy. In S. B. Hulley, S. R. Cummings, W. S. Browner, D. G. Grady, & T. B. Newman (Eds.), *Designing clinical research: An epidemiologic approach* (3rd ed., pp. 37–50). Philadelphia: Lippincott Williams & Wilkins.

Rodrigues, G. (2007). *Defining accuracy and precision* [Online]. Retrieved December 12, 2008, from http://www.encyclopedia.com/doc/1G1-168286146.html

Science 122 Laboratory. (2008). *The science of measurement: Accuracy vs. precision.* University of Hawaii. [Online]. Retrieved December 14, 2008, from http://honolulu.hawaii.edu/distance/sci122/SciLab/L5/accprec.html

# Predictive Validity

Validity refers to the degree to which a measure accurately measures the specific construct that it claims to be measuring. Criterion-related validity is concerned with the relationship between individuals' performance on two measures used to assess the same construct. It specifically measures how closely scores on a new measure are related to scores from an accepted criterion measure. There are two forms of criterion-related validity: predictive validity and concurrent validity.

Concurrent validity focuses on the extent to which scores on a new measure are related to scores from a criterion measure administered at the same point in time, whereas predictive validity uses the scores from the new measure to predict performance on a criterion measure administered at a later point in time.

Examples of contexts where predictive validity is relevant include the following:

- Scores on a foreign language aptitude measure given at the beginning of an immersion course are used to predict scores on a fluency exam administered at the end of the program.
- Scores on an employment measure administered to new employees at the time of hire are used to predict end-of-quarter job performance ratings from a supervisor.

The primary reason that predictive validity is of interest to users is that a concurrent criterion measure may not be available at the point in time at which decisions must be made. For example, it is not possible to evaluate a student's first-year college success at the time he or she is submitting college applications. Therefore, a measure that is able to correctly identify individuals who are likely to succeed in college at the time of application is a highly desirable tool for admissions counselors.

Before users can make decisions based on scores from a new measure designed to predict future outcomes reliably, they must have evidence that there is a strong relationship between the scores on the measure and the ultimate performance of interest. Such evidence can be obtained through a *predictive validation study.* In such a study, the new measure is administered to a sample of individuals

that is representative of the group for whom the measure is intended to be used. Next, researchers must allow enough time to pass for the behavior being predicted to occur. Once it has occurred, an already existing criterion measure is administered to the sample. The strength of the relationship between scores on the new measure and the scores on the criterion measure indicates the degree of predictive validity of the new measure.

The results of a predictive validation study are typically evaluated in one of two ways depending on the level of measurement of the scores from the two measures. In the case when both sets of scores are *continuous*, the degree of predictive validity is established via a *correlation coefficient*, usually the Pearson product-moment correlation coefficient. The correlation coefficient between the two sets of scores is also known as the *validity coefficient*. The validity coefficient can range from –1 to +1; large coefficients close to 1 in absolute value indicate high predictive validity of the new measure.

Figure 1 displays hypothetical results of a predictive validation study reflecting a validity coefficient of .93. The predictive validity of the aptitude measure is quite satisfactory because the aptitude measure scores correlate highly with the final exam scores collected at the end of the program; simply put, individuals scoring well on the aptitude measure later score well on the final exam.

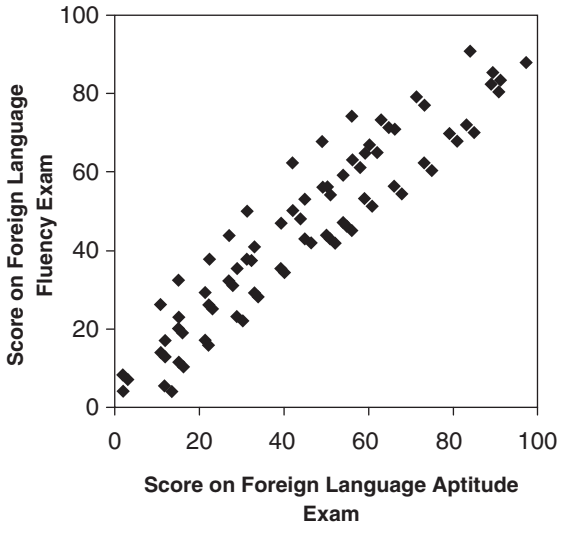



**Figure 1** Scatterplot of Scores on a Foreign Language Aptitude Exam and Language Fluency Exam

**Table 1** Cross Classification of Classification From an Employment Measure and Performance Rating

| *Initial Classification From Employment Measure* | *End-of-Quarter Job Performance Rating From Supervisor* | |
|---|---|---|
| | Unsatisfactory | Satisfactory |
| Unsatisfactory | 14 | 3 |
| Satisfactory | 7 | 86 |

In the case when the outcomes on both measures are classifications of individuals, *coefficients of classification agreement* are typically used, which are variations of correlation coefficients for *categorical data*. Evidence of high predictive validity is obtained when the classifications based on the new measure tend to agree with classifications based on the criterion measure.

Table 1 displays hypothetical data for a predictive validation study reflecting a classification consistency of 91%. The predictive validity of the employment measure is high because the resulting classification aligns well with the supervisor's job performance rating in almost all cases; simply put, the outcome of the measure predicts well the rating of the supervisor.

When determining the predictive validity of a new measure, the selection of a valid criterion measure is critical. Ideally, as noted by Robert M. Thorndike, criterion measures should be relevant to the desired decisions, free from bias, and reliable. In other words, they should already possess all the ideal measurement conditions that the new measure should possess also. Specifically, criterion measures should be

- *relevant to the desired decisions*—scores or classifications on the criterion measure should closely relate to, or represent, variation on the construct of interest. Previous validation studies and expert opinions should demonstrate the usefulness and appropriateness of the criterion for making inferences and decisions about the construct of interest.
- *free from bias*—scores or classifications on the criterion measure should be free from bias, meaning that they should not be influenced by anything other than the construct of interest.

Specifically, scores should not be affected by personal characteristics of the individual, subjective opinions of a rater, or other measurement conditions.
- *reliable*—scores or classifications on the criterion measure should be stable and replicable. That is, conclusions drawn about the construct of interest should not be clouded by inconsistent results across repeated administrations, alternative forms, or a lack of internal consistency of the criterion measure.

If the criterion against which the new measure is compared is invalid because it fails to meet these quality standards, the results of a predictive validation study will be compromised. Put differently, the results from a predictive validation study are only as useful as the quality of the criterion measure that is used in it. It is thus key to select the criterion measure properly to ensure that a lack of a relationship between the scores from the new measure and the scores from the criterion measure is truly due to problems with the new measure and not due to problems with the criterion measure.

Of course, the selection of an appropriate criterion measure will also be influenced by the availability or cost of the measure. Thus, the practical limitations associated with criterion measures that are inconvenient, expensive, or highly impractical to obtain may outweigh other desirable qualities of these measures.

*Jessica Lynn Mislevy and André A. Rupp*

***See also*** Concurrent Validity

**Further Readings**

Crocker, L., & Algina, J. (1986). *Introduction to classical & modern test theory.* Belmont, CA: Wadsworth.
Thorndike, R. M. (2005). *Measurement and evaluation in psychology and education* (7th ed.). Upper Saddle River, NJ: Pearson.

# Predictor Variable

Predictor variable is the name given to an independent variable used in regression analyses. The predictor variable provides information on an associated dependent variable regarding a particular outcome. The term *predictor variable* arises from an area of applied mathematic that uses probability theory to estimate future occurrences of an event based on collected quantitative evidence.

Predicted outcomes have become part of colloquial phrases in modern language. People speak about having higher risks associated with certain lifestyle choices, or about students' predicted academic performance in a university setting based on scores from standardized assessments. Examples abound in common language regarding predictive variables, yet confusion and fundamental misunderstandings surround the topic. Because of the frequent use and communication of predictor variables in behavioral and medical sciences, it is important for researchers to understand and clearly communicate information related to this topic.

This entry addresses basic assumptions associated with predictor variables, and the basic statistical fomula used in linear regression. It begins with an explanation of basic concepts and key words commonly associated with predictor variables; following this, basic linear regression fomulas are subjected to analysis with an emphasis on predictor variables. The entry concludes with a familiar example of predictor variables used in common applications.

## Basic Concepts and Terms

At the most fundamental level, predictor variables are variables that are linked with particular outcomes. As such, predictor variables are extensions of correlational statistics. Therefore, it is important to understand basic correlational concepts. As a function of the correlational relationship, when strength increases, slope, directional trends, and clustering effects become increasingly defined. Slope defines the degree of relationship between variables, with stronger relationships evidencing a steeper slope. The degree of slope varies depending on the relationship of the examined variables. Directional trends are either positive—as the $x$ variable increases, the $y$ variable increases—or negative—as either the $x$ or the $y$ variable increases, the other variable decreases. An example of a positive relationship is height and weight. As

people grow, weight typically increases. The relationship is not necessarily a one-to-one ratio, and exceptions certainly exist; however, this relationship is frequently observed. The relationship between weight of clothing worn by people and temperature is an example of a negative relationship between variables. People are unlikely to wear fewer clothing items on days when the weather is below the freezing point, and they are just as unlikely to wear layers of clothing when the temperature nears 100° Fahrenheit. As mentioned in the previous example, variation will exist in the population—some will choose to wear a light jacket on warm days, whereas others may wear T-shirts and shorts. This variation is called *clustering*. Correlational clusters are the degree to which variables form near each other around an estimate line that provides a representation of a linear trend in the scatterplot. Researchers commonly refer to this linear estimate as the line of best fit. As the observed spread in the cluster decreases, the observed variables exhibit less variance from the line of best fit; this increases the precision of the linear estimate. In a perfect correlational relationship, no deviations from the line of best fit occur, although this is a rare phenomenon.

Traditional correlational and regression models represent the dependent variable—or the variable the researcher is trying to influence—as "*y*." The predictor variable, "*x*," is the item that serves as the independent variable. Once a relationship between a dependent and an independent variable is established, and basic statistical assumptions are met, regression models expand on correlational assumptions. In regression analysis, however, the researcher's primary interest no longer lies in examining the relationship between two or more variables; instead, the researcher's focus turns to predicting a dependent variable ($y$) based on information obtained about an independent ($x$) variable. Like the correlational model, information regarding the line of best fit, clustering, direction, and slope is still needed. In regression applications, the line of best fit is called the *regression line,* and the analysis of clustering around the line of best fit becomes a fomulaic statistical procedure referred to as the *standard error of estimate*. The standard error of estimate defines the range of error associated with a prediction. For example, for every $x$ variable, the regression line identifies a predicted $y$ score; although a formal prediction is made, there is a range of scores associated with the $x$ score that deviates from the predicted score and leaves the researcher with error in his or her prediction. As an example, imagine a measure that claims to predict an outcome of interest. A score of 10 on this particular test predicts an outcome of 100 on the outcome measure. However, the standard error of estimate in this predictive equation is $x$ plus or minus 5 points. Therefore, the range of error associated with the outcome measure assumes that a subject obtaining a score of 10 on the independent—or predictive—test will likely obtain a score within 95–105 on the outcome measure. This error reveals that although the predictor variable is useful, it (like most predictive equations) remains imperfect.

Another important concept in predictive variables includes the idea of multiple predictor variables. Assuming that the research design and data collection methods meet sample size and other necessary statistical assumptions, many researchers often prefer to include multiple predictors over one predictor in an attempt to improve the precision and strength of the predictive equation. However, it is important for researchers to recognize that many variables share common variance among them. This simply means that the factors measured in one variable are also measured in another variable, either directly or indirectly, providing a form of statistical redundancy. For example, including height and weight as two predictors may provide an inflated estimate of the predictive accuracy of an equation because, as previously mentioned, when height increases, so does weight. In a situation where weight provides most of a researcher's predictive information, including height might prove useless because it provides another proxy of weight; therefore, it is as if the researcher counted weight twice.

## Predictor Variables in a Regression Fomula

Statistical textbooks present the standard regression equation in two forms. In secondary education settings, the equation is often expressed as $y = mx + b$. Where $y$ represents the predicted variable, $m$ refers to the slope of the line, $x$ represents the predictor variable, and $b$ is the point at which the regression line intercepts with the

*Y* axis. Although this fomula is popular in secondary education textbooks, it is expressed differently among postsecondary-level statistic books, where the fomula is written as $\hat{y} = a + bX$. Despite the various representations of this fomula, the integral ideas behind it remain intact. The meanings of the symbols in this fomula have changed; now, the $\hat{y}$ (read "y-hat") represents the predicted value of $y$ in the previous fomula, $a$ in the new equation is the $Y$ intercept, $b$ represents the slope of the regression line, and $x$ remains the dependent variable.

The linear regression equation in its current form allows one predictor variable, yet multiple predictor variables are easily included in regression equations with simple modifications. Using multiple predictors in one equation, researchers are able to use additional information when useful. It is important to note, however, that by using multiple predictive variables, it is not a simple matter of adding the two predictive variables together to create a superior predictive model. When multiple predictive variables are used, they are bound to have some overlap of common variance previously discussed, and researchers or statisticians should remain cognizant of potential errors.

The discussion to this point has centered on predictive equations with linear relationships using continuous predictive variables through simple regression models. However, other methods of statistical prediction are available. Other methods allow researchers to use dichotomous dependent or independent variables, examine curvilinear relationships, or examine classification error rates. For example, it is possible to use categorical independent variables in linear regression to make predictions of outcomes through procedures such as dummy coding or effect coding. In dummy coding, dichotomous variables are recorded through their occurrence or nonoccurrence. Researchers code variables into zeros or ones to represent the occurrence of an event, where zeros represent the nonoccurrence of an event and ones represent an occurrence. Dummy coding is often used to make predictions regarding categorical variables such as sex, ethnicity, or other demographic information. Similar to dummy coding, effect coding uses ones and zeros with the addition of negative and positive signs added to the one values. Effect coding is often used to differentiate between treatment effects.

When the dependent variable is a categorical metric, methods such as discriminate or logistic regression are more useful. Discriminate regression methods and logistic regression methods are similar in that they allow researchers to measure dichotomous outcome variables (e.g., the occurrence or nonoccurrence of an outcome). Logistic regression methods also allow researchers to determine classification rates of variables of interest to determine the accuracy of a measure. For example, researchers use logistic regression methods to determine correctly and incorrectly identified subjects in order to examine the precision of diagnostic tools.

## Practical Example

Now let us consider an applied example of predictive variables. Imagine that you work for a credit card company. Your company uses a widely known credit rating system to determine the appropriate credit limit and interest rate to assign to a customer. However, you realize that the credit system is based upon general recommendations of the major credit score companies. These scores do not account for your company's specialized customer base, which differs from mainstream customers typically represented in research studies. Interested in improving your company's prediction of creditworthiness, you decide to review your company's records of past clients to determine if your predictions are similar to or better than those recommended by the credit score companies. In this example, credit scores will serve as predictive or independent variables, and credit outcomes will serve as dependent variables.

Upon accessing your company's records, immediately you realize that there may be a problem with your research plan. Because your company is small, it cannot absorb high-risk failures; therefore, the company uses the most conservative estimate of creditworthiness based upon three available credit scores. Again, your company does this to avoid high-risk customers who might default on their credit card debt. In your analysis, you are confronted with a major methodological question. Should you continue to use the most conservative estimate and possibly deny potential

credit-worthy customers? Perhaps in your company's business model, this approach is the most intuitive; however, this would require three separate statistical analyses resulting in three separate prediction equations. Concerned that the complexity of this approach might cause more selection errors in application due to employees' using the wrong statistical fomula, you choose instead to use a statistical analysis method that will yield one fomula for your company's employees to use.

Next, you consider your fomulaic method, as always with the intent to provide the best predictive equation. You are considering using one of the following three approaches: a weight of the three scores in one formal equation, an average of the three credit-ranking scores, or three separate analyses of the three credit-ranking scores to determine which company provides the most precise information about your customer base. If you use a weighted approach that includes all three scores, your analysis will provide an inflated predictive estimate. Said differently, your predictive equation will look as if it is stronger than it actually is. This is because these three credit-ranking scores measure many of the same customer characteristics that the other credit-score companies measure (e.g., the frequency that customers pay their bills on time, the amount of debt they have, if they have previously defaulted on other debts). These overlapping measures provide statistical redundancies. Realizing that you wish to avoid unnecessary error, you decide that the amount of shared variance among the three scores is too high and will likely provide false confidence in your analysis.

You then consider using an arithmetic average of the three credit score rankings. However, because of the sometimes large variability among credit rating scores, as well as incomplete information, you determine that averaging the three scores may result in a less precise measure of a credit ranking. Therefore, you consider analyzing the three credit scores individually and using the scores from the credit ranking company that provides the strongest predictive equation. This approach is perhaps the most intuitive. Your analysis will reveal the most precise information for your customer base while reducing unnecessary fees from redundant and less precise information provided by other companies. Additionally, because only one score is presented to your customer service employees—from which they will make their decisions—fewer errors in the approval process are likely.

*Justin P. Allen*

***See also*** Correlation; Discriminant Analysis; Dummy Coding; Effect Coding; General Linear Model; Logistic Regression; Raw Scores; Scatterplot; Standard Deviation; Structural Equation Modeling; Variance

**Further Readings**

Green, S. B., & Salkind, N. J. (2007). *Using SPSS for Windows: Analyzing and understanding data* (5th ed.). Englewood Cliffs, NJ: Prentice Hall.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction* (3rd ed.). Toronto: Wadsworth.

Serber, G. A. F., & Lee, A. J. (2003). *Linear regression analysis* (2nd ed.). Hoboken, NJ: Wiley.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston: Pearson.

## Pre-Experimental Designs

In pre-experimental designs, either a single group of participants or multiple groups are observed after some intervention or treatment presumed to cause change. Although they do follow some basic steps used in experiments, pre-experimental designs either fail to include a pretest, a control or comparison group, or both; in addition, no randomization procedures are used to control for extraneous variables. Thus, they are considered "pre-," indicating they are preparatory or prerequisite to true experimental designs. Pre-experimental designs represent the simplest form of research designs. Together with quasi-experimental designs and true experimental (also called randomized experimental) designs, they make the three basic categories of designs with an intervention. Each contains subdesigns with specific strengths and weaknesses.

Because the validity of pre-experimental designs is threatened by inadequate control during implementation, it is difficult or impossible to rule out rival hypotheses or explanations. Therefore, researchers should be especially cautious when

interpreting and generalizing the results from pre-experimental studies. However, pre-experimental designs are cost-effective ways to explore whether a potential intervention merits further investigation. They might be particularly useful when there are less than perfect conditions for true experimental designs. These may include the restraints of time, space, participants, resources, and ethical issues; logistical constraints, such as the investigation of a small atypical sample; or the evaluation of a previous intervention without adequate planning of the research design. Under those circumstances, as long as appropriate caution is taken while making interpretations and generalizations, results from pre-experimental designs could still be suggestive to the field and future research.

This entry first describes the notations used to diagram pre-experimental designs. Next, this entry discusses the major types of pre-experimental designs and possible improvements to these designs. The entry concludes with a brief presentation of the limitations and benefits of pre-experimental designs.

## Diagramming Pre-Experimental Designs

Some useful designations are widely used to clarify the different designs. NR indicates that the assignment of participants into groups is nonrandom. E represents the group of participants exposed to the experimental variable or event (usually an intervention or treatment), whereas C represents the control or comparison group. The intervention or treatment is expressed as X, the effects of which are to be measured, whereas ~X either means no intervention or the usual treatment. Thus, X and ~X signify two different levels of the independent variable in a study. O refers to the observation or measurement of the dependent variable. This can be done by observing behaviors, conducting interviews, administering tests, and so on. The X (or ~X) and Os in a given row are applied to the same group of participants, either E or C. The left-to-right dimension reflects the temporal order.

## Types of Pre-Experimental Designs

Three major types of pre-experimental designs are commonly used, either because the research preplanning is inadequate, causing unanticipated problems, or perhaps the situation makes a pretest or a comparison group impossible.

### *One-Group Posttest-Only Design*

The first pre-experimental design, the one-group posttest-only design, is also called the one-shot case study or case study design. In this subdesign, all the participants are assigned nonrandomly into the intervention group. Then, this whole group of participants is presented with some kind of intervention, such as a new curriculum, a new drug, or a new service. After that, the outcome performance is measured as the only posttest. The design is shown as follows:

| Group | Intervention | Posttest |
|---|---|---|
| E : | X | O |

Although the goal of this design is to justify the effect of the intervention on the outcome, it is impossible without a pretest to know whether any change has occurred to this group. It is also impossible to determine whether the intervention has made a difference compared to the situation where no intervention has been implemented because there is no comparison group. Thus, such a design does not satisfy even the minimum condition for a research problem to investigate a relationship or comparison. Therefore, it does not merit the title of experiment.

Obviously, the validity of this design is seriously threatened in many ways. First, the participants are not randomly selected, so external validity could not be determined. Second, the participants are not randomly assigned into experimental and comparison groups, and no steps are taken to ensure that the intervention was the only thing going on during the process. Historical events or maturation may also have an influence on the posttest results. But nothing else is measured to determine what extraneous factors may have been confounded with the intervention. Thus, internal validity is threatened.

With only the posttest scores in hand, this one-group posttest-only design merely provides a reference point. If other specific background information is available, such as the results from an earlier group, an earlier time point, or the general population, the results may indicate the intervention's

effect to some extent and be suggestive for future study.

### *One-Group Pretest–Posttest Design*

The second pre-experimental design is the one-group pretest–posttest design. In this approach, all the participants are first assigned to the experimental group. Then, the group is observed at two time points. One observation in the form of a pretest is recorded before the intervention (independent variable). And then after the intervention, a second observation is conducted in the form of a posttest. Changes from the pretest to the posttest are the outcomes of interest, which are presumed to be the result of the intervention. However, there is no comparison group. The design could be shown as follows:

$$E: \quad O_1 \quad X \quad O_2$$

This pre-experimental design is superior to the previous one in that it has included a pretest to determine the baseline score. This makes it possible to compare the group's performance before and after the intervention. So now, researchers at least know whether a change in the outcome or dependent variable has taken place within the same group. But without a comparison group, it is still impossible to clarify whether this change would happen even without the application of the intervention; that is, whether any change from the pretest to the posttest is due to the intervention only and not due to other extraneous variables.

Several categories of uncontrolled extraneous variables could lead to alternative explanations of any differences between the pretest and the posttest, confusing the possible effects of the intervention.

The first extraneous variable is related to history. That is, during the time span between the pretest and the posttest, many other events may have occurred in addition to the intervention. Thus, the results could be attributed to these as well as to the intervention. In the second place, maturation changes in the participants could also produce differences between the pretest and the posttest scores. As people get older, they may become better at the outcome variable at the same time as the intervention occurs. Similarly, carryover effects are also a possible problem in this design because taking the pretest could influence the posttest. Participants may have learned from the pretest and perform better at the posttest if the two measurements are the same or overlap a lot. Their performance may also vary because of the difference in either the instruments or the implementation processes between the two. Finally, if the experimental group is selected because of extreme scores on some type of instrument, such as the very best or the very worst on a test, then the differences between the pretest and the posttest scores could be due to statistical regression toward the mean. There is a tendency for extreme scorers to move toward the average or more typical performance on subsequent tests.

Any of these above factors may cast threats on the validity of this design. The failure to rule out plausible alternative explanations makes it very difficult to generalize with any kind of confidence about any causal relation between the intervention and the observed change.

However, the one-group pretest–posttest design still allows ruling out some alternative hypotheses by including both pretest and posttest and offering suggestive implications for future study.

### *Posttest-Only Nonequivalent Groups Design*

The third pre-experimental design is the posttest-only nonequivalent groups design, which is also called static-group comparison or cross-sectional study. In this subdesign, participants are assigned into two groups for comparison. However, the experimental and comparison groups again are determined nonrandomly. After that, the experimental group receives the treatment and the comparison group does not. The subsequent difference between the two groups, which is assumed to reflect the influence of the treatment, is measured by the posttest. This design could be thought of as adding a comparison group to the one-group posttest-only design. It could be diagrammed as follows:

$$\begin{array}{llll} NR & E: & X & O \\ NR & C: & \sim X & O \end{array}$$

This design attempts to make a comparison by including a group that does not receive the

treatment. However, without randomization and a pretest, it is difficult to know whether there is any pre-existing difference between the two groups. The nonequivalent group is picked up for the purpose of comparison, but its comparability is questionable. Any differences between the two groups before the treatment could become rival explanations for the difference at posttest. It would be difficult to make a conclusion about the effects of the treatment because of unknown differences between the groups. Ideally, with a comparison group, one could be more confident about the intervention's effect if the two groups do go through everything in the same way from the same environment, except for the treatment. However, without any assurance about the equality of the two groups, extraneous factors such as history, maturation, attrition, and statistical regression may affect the two groups differently and contaminate the result at the posttest. An example of history could be that participants in the comparison group seek improvement from some other programs or, on the other hand, become discouraged by the design. Also, if the two groups are nonequivalent at the beginning, the maturation process might be very different across the same time span. If more participants in the experimental group drop out from the treatment, or more participants in the comparison group drop out because they do not feel that they have gained any benefits, such unequal attrition could become a serious threat to the validity of the design. As for statistical regression, it could always be a threat if one or both of the groups were selected from an extreme population.

This pre-experimental design offers us more information and confidence than the one-group pretest-only or the one-group pretest–posttest design by including a comparison group, as long as the above problems related to nonrandom assignment and lack of pretests are well considered when making explanations and implications.

## Improvements to Pre-Experimental Designs

### *One-Group Posttest-Only Design*

By following Donald Campbell's reasoning about pattern matching, the one-group posttest-only design could be improved in interpretability. Such improvement is achieved through adding multiple, unique, and substantive posttests to the original design. It could be expressed as follows:

$$E: \quad X \quad O_{1A} \quad O_{1B} \quad \cdots \quad O_{1N}$$

Here, the posttests ($O_{1A}$, $O_{1B} \cdots O_{1N}$) represent measurements of different constructs after the intervention. Theoretically, when the possible causes are known, the pattern of effects could be matched by the results from the measurements. Suppose a specific and observable pattern of results is predicted by the implementation of an intervention. The researcher could conduct multiple posttests to collect data about each of the results so that the effects of the intervention could be either supported or disputed. Such collection of convergent evidence is different from the previous design, in which only one posttest construct is assessed.

However, the pattern-matching logic is most compelling when the effects are known and the causes are to be discovered retrospectively, such as the detection of a series of clues left by a criminal. But the pre-experimental designs work just the opposite way. In a prospective study, the causes are predetermined and the effects are unknown. To decrease the possibility of Type I errors, which are very likely considering the human tendency to spot patterns even in random data, well-reasoned specification of a unique pattern beforehand is required to support the presumed relation between intervention and its effects.

### *One-Group Pretest–Posttest Design*

Some variations could also be made to improve the interpretability of the one-group pretest–posttest design. The threats from maturation and regression could be reduced by simply adding a second pretest prior to the first pretest, which could be represented as follows:

$$E: \quad O_1 \quad O_2 \quad X \quad O_3$$

Preferably, the time between $O_1$ and $O_2$ equals the time between $O_2$ and $O_3$. With two pretests, the potential biases that might exist in the difference between the second pretest and the posttest, such as the natural improvement in performance as the participants mature or the tendency to regress to the average level, could be clarified by

investigating the difference between the two pretests. If the difference between $O_2$ and $O_3$ is much greater than the difference between $O_1$ and $O_2$, one could be pretty confident in stating that the intervention (X) has made a difference. However, in cases where the trend of maturation is nonlinear, more pretests before the intervention would be needed to capture that.

Similar to the improvement to the one-group posttest-only design discussed earlier, adding another unique and substantive dependent variable to the original one-group pretest–posttest design could also help reduce potential threats. This could be diagrammed as follows:

$$E: \quad O_{1A} \quad O_{1B} \quad X \quad O_{2A} \quad O_{2B}$$

Here, in addition to Measure A, which is assessed during both pretest and posttest, a nonequivalent Measure B is also assessed at the same two time points to the same experimental group. If Measure A is hypothesized to change with the implementation of the intervention and Measure B is expected to remain the same, both of the measures would face the validity threats in the same way. With such divergent evidence between two mutually exclusive dependent variables, one could be much more confident about the effects of the intervention because both of the dependent variables are exposed to the same set of environment causes to the same degree.

Another improvement could be made by converting the design into a time series design as follows:

$$E: \quad O_1 \quad O_2 \quad O_3 \quad O_4 \quad X \quad O_5 \quad O_6 \quad O_7 \quad O_8$$

In such a design, several earlier observations (Pretests 1, 2, and 3) and several later observations (Posttests 6, 7, and 8) are added. Suppose there was little change observed between Pretest 1, Pretest 2, and Pretest 3 prior to the intervention, but observable changes occurred after the intervention in the posttest, it would be pretty convincing to state that the intervention or treatment has altered the trend of target development. The key point here is to have multiple (at least three) pretests to establish a baseline.

### *Posttest-Only Nonequivalent Groups Design*

Because the major weakness of the posttest-only nonequivalent groups design is the nonequivalence of the two groups, one possible improvement is to have a pretest on the dependent variable for both groups. Then the design changes to a better, but still weak, pretest–posttest nonequivalent comparison group design. Even if the groups were equivalent on the dependent variable at the pretest, they may still differ on other variables such as demographics. Therefore, other statistical attempts could be applied to check whether the groups differ on demographic or other available information. If no statistically significant differences could be found on gender, ethnicity, social class, and so on between the two groups, one would be more confident to consider them similar. However, "not significantly different" does not ensure equivalence, and one can never measure all of the possible crucial participant characteristics in actual research.

Sometimes, it is impractical or even unethical to have a comparison group that does not receive the treatment. Under such circumstances, a wait-list comparison group design is often recommended. Here, all participants would receive the treatment eventually, but some are assigned, preferably randomly, to a waiting list that would receive the treatment later than the group with immediate intervention. Both groups are assessed when they first enter the study. The immediate intervention group would receive the treatment then, and the wait-list group would not. After that, a second assessment is conducted on both groups. Then it is the wait-list group's turn to receive the same treatment as the immediate intervention group received. This group would be assessed again after the treatment. The diagram of the wait-list comparison group design could be shown as follows:

| | | | | | |
|---|---|---|---|---|---|
| Immediate Intervention Group | $O_1$ | X | $O_2$ | | |
| Wait-List Group | $O_1$ | ~X | $O_2$ | X | $O_3$ |

Although such design gives all the participants opportunity for treatment, it is only practical when the intervention is relatively brief, a few months at

most, and when it is ethical and practical to have the wait-list group wait for its treatment.

### Limitations and Benefits

Pre-experimental designs do not have random assignment of participants and are missing a comparison group, a pretest, or both. Therefore, they provide very little support for the effectiveness of the intervention or treatment, but they are useful and flexible when limited resources or ethical issues make well-controlled experiments impossible. In spite of their limitations, pre-experimental designs could offer some basic descriptive information for further research.

*Jun Wang and George A. Morgan*

***See also*** Experimental Design; Nonexperimental Design; Pretest–Posttest Design; Quasi-Experimental Design; Single-Subject Design; Time-Series Study; True Experimental Design

#### Further Readings

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin, 54,* 297–311.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings.* Boston: Houghton Mifflin.

Gliner, J. A., Morgan, G. A., & Leech, N. L. (2009). *Research methods in applied settings: An integrated approach to design and analysis* (2nd ed.). New York: Routledge, Taylor and Francis Group.

Morgan, G. A., Gliner, J. A., & Harmon, R. J. (2006). *Understanding and evaluating research in applied and clinical settings.* Mahwah, NJ: Lawrence Erlbaum.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin.

Sheskin, D. (2000). *The handbook of parametric and nonparametric statistical procedures.* Boca Raton, FL: Chapman and Hall.

Slavin, R. E. (2007). *Educational research in an age of accountability.* Boston: Pearson Education.

White, R. T., & Arzi, H. J. (2005). Longitudinal studies: Designs, validity, practicality, and value. *Research in Science Education, 35,* 137–149.

## Pretest–Posttest Design

The basic premise behind the pretest–posttest design involves obtaining a pretest measure of the outcome of interest prior to administering some treatment, followed by a posttest on the same measure after treatment occurs. Pretest–posttest designs are employed in both experimental and quasi-experimental research and can be used with or without control groups. For example, quasi-experimental pretest–posttest designs may or may not include control groups, whereas experimental pretest–posttest designs must include control groups. Furthermore, despite the versatility of the pretest–posttest designs, in general, they still have limitations, including threats to internal validity. Although such threats are of particular concern for quasi-experimental pretest–posttest designs, experimental pretest–posttest designs also contain threats to internal validity.

### Types of Pretest–Posttest Designs Without Control Groups

#### *One-Group Pretest–Posttest Design*

In the simplest pretest–posttest design, researchers gather data about some outcome through a single pretest, administer a treatment, and then gather posttest data on the same measure. This design is typically represented as follows:

$$O_1 \; X \; O_2$$

where $O_1$ represents the pretest, X represents some treatment, and $O_2$ represents the posttest.

Including a pretest measure is an improvement over the posttest-only research design; however, this design is still relatively weak in terms of internal validity. Although this design allows researchers to examine some outcome of interest prior to some treatment ($O_1$), it does not eliminate the possibility that $O_2$ might have occurred regardless of the treatment. For example, threats to internal validity, such as maturation, history, and testing, could be responsible for any observed difference between the pretest and posttest. Also, the longer the time lapse between the pretest and posttest, the

harder it is to rule out alternative explanations for any observed differences.

When the outcome of interest is continuous, data obtained from the one-group pretest–posttest design can be analyzed with the dependent-means *t* test (also sometimes called the correlated-means *t* test or the paired-difference *t* test), which tests $H_0 : \mu_D = 0$. When the outcome of interest is categorical and has only two levels, one-group pretest–posttest data can be analyzed with McNemar's chi-square to test the null hypothesis that the distribution of responses is the same across time periods. However, when the outcome of interest is categorical and has more than two levels, McNemar's chi-square cannot be used. Instead, data can be analyzed through Mantel–Haenszel methods.

### *One-Group Pretest–Posttest Design Using a Double Pretest*

The one-group pretest–posttest design can be improved upon by adding a second pretest prior to treatment administration:

$$O_1 \ O_2 \ X \ O_3,$$

where $O_1$ and $O_2$ represent the two pretests, X represents some treatment, and $O_3$ represents the posttest.

Adding a second pretest to the traditional one-group pretest–posttest design can help reduce maturation and regression to the mean threats as plausible explanations for any observed differences. For example, instead of comparing $O_3$ only to $O_1$ or $O_2$, any observed difference between $O_2$ and $O_3$ can also be compared to any differences between $O_1$ and $O_2$. If the difference between $O_2$ and $O_3$ is larger than the difference between $O_1$ and $O_2$, then the observed change is less likely solely due to maturation.

When the outcome of interest is continuous, data obtained from the one-group pretest–posttest design using a double pretest can be analyzed through a within-subject design analysis of variance (ANOVA) (also sometimes referred to as repeated measures ANOVA) with appropriate contrasts to test the null hypothesis of interest, $H_0 : \mu_{post} - \mu_{pre2} = \mu_{pre2} - \mu_{pre1}$. When the outcome of interest is categorical, data from this design can be analyzed with conditional logistic regression (also referred to as subject-specific models) to test the null hypothesis that the distribution of responses is the same across time periods.

### *One-Group Pretest–Posttest Design Using a Nonequivalent Dependent Variable*

The simple one-group pretest–posttest design also can be improved by including a nonequivalent dependent variable. For example, instead of obtaining pretest and posttest data on the outcome of interest, with this design, the researcher obtains pretest and posttest data on the outcome of interest (A) as well as on some other outcome (B) that is similar to the outcome of interest, but is not expected to change based on the treatment.

$$(O_{1A}, O_{1B}) \ X \ (O_{2A}, O_{2B})$$

In this design, $O_{1A}$ represents the pretest on the outcome of interest, $O_{1B}$ represents the pretest on the nonequivalent outcome, X represents some treatment, $O_{2A}$ represents the posttest on the outcome of interest, and $O_{2B}$ represents the posttest on the nonequivalent outcome. Including a nonequivalent dependent variable helps researchers assess any naturally occurring trends in the data, separate from the treatment effect. If the posttest difference for the outcome of interest ($O_{2A} - O_{1A}$) is larger than the posttest difference for the secondary outcome ($O_{2B} - O_{1B}$), the less likely the change is solely a result of maturation.

When the outcome of interest is continuous, data obtained from the one-group pretest–posttest design using a nonequivalent dependent variable can be analyzed through a within-subject ANOVA design with appropriate contrast statements to test the null hypothesis of interest, $H_0$: $\mu_{postA} - \mu_{preA} = \mu_{postB} - \mu_{preB}$. When the outcome of interest is categorical, data from this design can be analyzed with conditional logistic regression to test the null hypothesis that the distribution of responses is the same across time periods and across the two dependent variables (i.e., outcome of interest and the nonequivalent dependent variable).

### *Removed-Treatment Design*

This design is also an improvement to the one-group pretest–posttest design. By adding multiple

posttests and the removal of the treatment to the one-group pretest–posttest design, the removed-treatment design allows better control over the internal validity of a study. That is, by examining the outcome of interest after the treatment has been stopped or removed, researchers have another piece of information that allows for multiple comparisons and helps inform their conclusions. This design is typically represented as

$$O_1 \; X \; O_2 \; O_3 \; \not{X} \; O_4,$$

where $O_1$ represents the pretest on some outcome of interest, X represents some treatment, $O_2$ represents the first posttest on the outcome of interest, $O_3$ represents the second posttest on the same outcome of interest, $\not{X}$ represents treatment removal, and $O_4$ represents the third posttest on the outcome of interest.

With this design, if a treatment effect exists, the difference between $O_2$ and $O_1$ should be greater than the difference between $O_4$ and $O_3$. Also, examining changes between $O_3$ and $O_2$ allows the researcher to investigate possible trend effects after treatment. However, this theoretical pattern of change is often difficult to observe unless a treatment effect dissipates quickly upon treatment removal (which is often not the case in the social and behavioral sciences). Furthermore, because removal of certain treatments might be considered unethical, considerations need to be discussed before employing this pretest–posttest design.

When the outcome of interest is continuous, data obtained from the removed-treatment design can be analyzed through a within-subject ANOVA with appropriate contrast statements to test the null hypothesis of interest, $H_0 : \mu_{post1} - \mu_{pre} = \mu_{post3} - \mu_{post2}$. When the outcome of interest is categorical, data from this design can be analyzed with conditional logistic regression to test the null hypothesis that the distribution of responses is the same across time periods.

### *Repeated-Treatment Design*

Building upon the removed-treatment design, the repeated-treatment design includes multiple posttest observations, treatment removal, and treatment reintroduction. In the following notation, $O_1$ represents the pretest on some outcome of interest, X represents some treatment, $O_2$ represents the first posttest on the outcome of interest, $\not{X}$ represents treatment removal, $O_3$ represents the second posttest on the same outcome of interest, and O4 represents the third posttest on the outcome of interest.

$$O_1 \; X \; O_2 \; \not{X} \; O_3 \; X \; O_4$$

If a treatment effect is present, in this design, $O_2$ should differ from $O_1$ and from $O_3$; however, the difference between $O_3$ and $O_2$ should be opposite or less than the difference between $O_2$ and $O_1$. Similarly, $O_4$ should differ from $O_3$, and the difference between $O_4$ and $O_3$ should be similar to the difference between $O_2$ and $O_1$. Although the inclusion of a second treatment period adds rigor to this design, it is not without limitations. As with the removed-treatment design, this design is best when the treatment effect will truly dissipate upon removal of the treatment and when it is ethically acceptable to remove the treatment. Furthermore, by reintroducing the treatment, participants are more likely to become aware of the treatment and researchers' expectations, thus increasing the potential of contamination.

When the outcome of interest is continuous, data obtained from the repeated-treatment design can be analyzed through a within-subject design ANOVA with appropriate contrast statements to test the null hypotheses of interest, $H_0 : \mu_{post1} - \mu_{pre} = \mu_{post2} - \mu_{post1}$ and $H_0 : \mu_{post2} - \mu_{post1} = \mu_{post3} = \mu_{post2}$. When the outcome of interest is categorical, data from this design can be analyzed with conditional logistic regression to test the null hypothesis that the distribution of responses is the same across time periods.

## Types of Pretest–Posttest Designs With Control Groups

### *Two-Group Pretest–Posttest Design*

Similar to the one-group pretest–posttest design, this design includes a nontreated control group and is often represented as follows:

$$O_1 \; X \; O_2$$
$$O_1 \quad O_2,$$

where $O_1$ represents the pretest, X represents some treatment, and $O_2$ represents the posttest. When randomization is employed, the two-group pretest–posttest design becomes the classic experimental design.

Whether the two-group pretest–posttest design is used in experimental or quasi-experimental research, including an untreated control group reduces threats to internal validity and allows for within-group and between-group comparisons. For example, although selection bias is still a concern when this design is used in quasi-experimental research, use of a pretest and a comparison group allows researchers to examine the nature and extent of selection bias by comparing the treatment and control groups before the treatment is administered. Noting any pretest differences between the groups allows for stronger inferences to be made after the treatment is administered. Also, although threats such as maturation and history still might exist with the two-group pretest–posttest design, the effects of these threats should be the same for both groups, thus adding more support for a treatment effect when observed posttest differences exist between the groups.

When the outcome of interest is continuous, data obtained from the two-group pretest–posttest design can be analyzed with a one-between one-within ANOVA design (also referred to as a factorial repeated measures ANOVA) to test $H_0 : \mu_{\text{post_Tx}} - \mu_{\text{pre-_Tx}} = \mu_{\text{post_C}} - \mu_{\text{pre_C}}$. When the outcome of interest is categorical, two-group pretest–posttest data can be analyzed with generalized estimating equations (GEEs) to test the null-hypothesis that the distribution of responses is the same across time periods and between groups.

### *Two-Group Pretest–Posttest Design With a Double Pretest*

Improving upon the two-group pretest–posttest design, the two-group pretest–posttest design with a double pretest includes administering two pretests prior to treatment followed by one posttest measure after the treatment period.

$$\begin{array}{llll} O_1 & O_2 & X & O_3 \\ O_1 & O_2 & & O_3 \end{array}$$

In this design, $O_1$ represents the first pretest, $O_2$ represents the second pretest, X represents some treatment, and $O_3$ represents the posttest. One advantage of the two-group pretest–posttest design with a double pretest over the two-group pretest–posttest design is the ability to examine trends and selection bias in the treatment and control groups prior to treatment administration. If the treatment has an effect, the change between $O_3$ and $O_2$ for the treatment group should be different from the observed change during the same time period for the control group. This design also reduces threats to internal validity. By including the untreated control group, threats such as maturation, history, and regression to the mean would occur in both groups, thus, in essence, balancing the groups in terms of threats. Therefore, observed pretest–posttest treatment group differences are more likely to represent treatment effects.

When the outcome of interest is continuous, data obtained from the two-group pretest–posttest design with a double pretest can be analyzed with a one-between one-within ANOVA design with appropriate contrast statements to test the null hypothesis of interest, $H_0 : \mu_{\text{post_Tx}} - \mu_{\text{pre2-_Tx}} = \mu_{\text{post_C}} - \mu_{\text{pre2_C}}$. When the outcome of interest is categorical, data from this design can be analyzed with GEEs with appropriate contrast statements to test the null hypothesis, $H_0 : \mu_{\text{post_Tx}} - \mu_{\text{pre2-_Tx}} = \mu_{\text{post_C}} - \mu_{\text{pre2_C}}$.

### *Four-Group Design With Pretest–Posttest and Posttest–Only Groups*

Also known as the Solomon four-group design, this design is a combination of the two-group pretest–posttest design and the two-group posttest-only design. Pretest measures are obtained from the two pretest–posttest groups, treatment is administered to one of the pretest–posttest groups and to one of the posttest-only groups, and posttest measures are obtained from all four groups.

$$\begin{array}{llll} O_1 & X & O_2 & \\ O_1 & & O_2 & \\ & X & O_2 & \\ & & O_2 & \end{array}$$

In this design, $O_1$ represents the pretest, X represents some treatment, and $O_2$ represents the posttest.

Although this design is a bit more complicated than the simple two-group pretest–posttest design, the inclusion of the two posttest–only groups allows researchers to investigate possible testing threats to internal validity. For example, if $O_2$ for the treatment pretest–posttest group is similar to $O_2$ for the treatment posttest-only group, then testing effects are likely not present. Similarly, as with other multiple-group pretest–posttest designs, this design allows for both within-group and between-group comparisons, including examination of possible selection bias when this design is used in quasi-experimental research. With this design, maturation and history threats to internal validity are also diminished. If a treatment effect is present, then the difference between $O_2$ and $O_1$ should be the same for the treatment and control pretest-posttest groups, and $O_2$ for the two treatment groups should be similar to one another and different from $O_2$ for the two control groups.

When the outcome of interest is continuous, data obtained from the Solomon four-group design can be analyzed with two different statistical procedures. First, a one-between one-within ANOVA design can be used to test the null hypothesis $H_0 : \mu_{post_Tx} - \mu_{pre_Tx} = \mu_{post_C} - \mu_{pre_;C}$ and an independent means $t$ test can be used to test the null hypothesis $H_0 : \mu_{post_Tx2} - \mu_{post_C2}$. When the outcome of interest is categorical, data can be analyzed with GEEs to test analogous null hypotheses: $H_0 : \mu_{post_Tx} - \mu_{pre_Tx} = \mu_{post_C} - \mu_{pre_C}$ and $H_0 : \mu_{post_Tx2} - \mu_{post_C2}$.

### *Switching Replications Design*

The switching replications design refers to the process of obtaining pretest data from two groups, administering some treatment to one group, obtaining a second assessment measure from both groups, administering the same treatment to the second group, and obtaining a third assessment measure from both groups. This design is typically denoted as

$$\begin{array}{llll} O_1 & X & O_2 & O_3 \\ O_1 & & O_2 & X\ O_3, \end{array}$$

where $O_1$ refers to the pretest assessment for both groups, X represents some treatment, $O_2$ represents the first posttest for the first group and the second pretest for the second group, and $O_3$ represents a second posttest for the first group and the first posttest for the second group.

As with the other pretest–posttest designs that use a control group, if participants are not randomly assigned to the groups, selection bias is a likely threat. However, because pretest measures are obtained from both groups, potential group differences can be examined before the first treatment is administered. Furthermore, administering the treatment to both groups at different times helps improve the internal validity of the study. For example, if some trend were responsible for any observed difference between $O_2$ and $O_1$ for the first group, then one also would expect to see the same level of change during this time period in the second group. This same principle applies to differences between $O_3$ and $O_2$. Thus, if a treatment effect exists, the difference between $O_2$ and $O_1$ for the first group should be different from the difference between $O_2$ and $O_1$ for the second group, yet similar to the difference between $O_3$ and $O_2$ for the second group. However, as with the four-group design with pretest–posttest and posttest-only groups, the timing difference in treatment administration is an important issue that needs to be acknowledged when employing this design. The treatment administration to the second group can never be exactly the same as the treatment that is administered to the first group.

When the outcome of interest is continuous, data obtained from the switching replications design can be analyzed with a one-between one-within ANOVA design with appropriate contrast statements to test the null hypotheses of interest, $H_0 : \mu_{post1_Tx} - \mu_{pre_Tx} = \mu_{pre2_C} - \mu_{pre1_C}$ and $H_0 : \mu_{post2_Tx} - \mu_{post1_Tx} = \mu_{post_C} - \mu_{pre2_C}$. When the outcome of interest is categorical, data from this design can be analyzed using GEEs and analogous contrasts as those above, to test the null hypotheses of interest, $H_0 : \mu_{post1_Tx} - \mu_{pre_Tx} = \mu_{pre2_C} - \mu_{pre1_C}$ and $H_0 : \mu_{post2_Tx} - \mu_{post1_Tx} = \mu_{post_C} - \mu_{pre2_C}$.

### *Reversed-Treatment Pretest–Posttest Control Group Design*

With this design, one group receives the treatment of interest, expected to affect the outcome of

interest in one direction, and the other group receives an opposite treatment, expected to affect the outcome of interest in the opposite direction. Diagrammed as

$$O_1 \; X_+ \; O_2$$
$$O_1 \; X_- \; O_2,$$

$O_1$ represents the pretest, $X_+$ represents the treatment of interest, $X_-$ represents the treatment expected to produce opposite effects, and $O_2$ represents the posttest.

The basic premise behind this design is that if the difference between $O_2$ and $O_1$ for the group administered $X_+$ is in one direction and the difference between $O_2$ and $O_1$ for the group administered $X_-$ is in the opposite direction, a statistical interaction, suggesting a treatment effect, should be present. However, as with the other pretest-posttest designs that use a control group, if participants are not randomly assigned to the groups, selection bias is a likely threat and should be examined. Also, as with pretest–posttest designs that include treatment removal or delayed treatment administration, ethical issues often surround use of the reversed-treatment pretest–posttest control group design. Finally, conceptual difficulties are inherent in such a design as researchers must identify treatments that are expected to produce opposite effects on the outcome variable.

When the outcome of interest is continuous, data obtained from the switching replications design can be analyzed with a one-between one-within ANOVA design to test the null hypothesis of interest, $H_0: \mu_{\text{post_Tx}} - \mu_{\text{pre_Tx}} = \mu_{\text{post_C}} = \mu_{\text{pre_C}}$. When the outcome of interest is categorical, data can be analyzed with GEEs to test the null hypothesis that the distribution of responses is the same across time periods and between groups.

*Bethany A. Bell*

***See also*** Experimental Design; General Linear Model; Logistic Regression; Quasi-Experimental Design; Repeated Measures Design; Threats to Validity; Within-Subjects Design

**Further Readings**

Brogan, D. R., & Kutner, M. H. (1980). Comparative analyses of pretest-posttest research design. *American Statistician, 34,* 229–232.

Feldman, H. A., McKinlay, S. M. (1994). Cohort versus cross-sectional design in large field trials: Precision, sample size, and a unifying model. *Statistics in Medicine, 13,* 61–78.

Shadish, W. R., Cook, T. D., Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin.

Stokes, M. E., Davis, C. S., Koch, G. G. (2006). *Categorical data analysis using the SAS system.* Cary, NC: SAS Institute.

# Pretest Sensitization

Pretest sensitization refers to the potential or actuality of a pretreatment assessment's effect on subjects in an experiment. The term has different meanings in various disciplines. In medicine and biology, it typically refers to a physical reaction to an initial administration of a drug or regimen. In the social sciences, it has come to mean a cognitive or psychological change in a subject due to administration of a test or observation of the subject. Although the mechanism of sensitization is typically not investigated or even known, the effect on behavior or response has been long documented. This is referred to here as the *pretest effect.* This entry focuses on the research of this phenomenon, methods of evaluation, and related statistical considerations.

## Research

Pretest effects have been of concern for many decades. Glenn H. Bracht and Gene V Glass discussed pretest effects in the context of external validity problems in interpreting research. That is, if a pretest is given in the context of the experiment, but will not normally be employed in practice, what is the potential effect of the pretest on the outcomes, because in effect it has become part of the treatment? Few authors have followed up this line of critical analysis of research in the intervening

years, so this external validity problem is probably understudied and underreported.

Victor L. Willson and Richard R. Putnam conducted a meta-analysis of randomized experiments that directly examined the effects of giving a pretest versus not giving one for various kinds of tests, including cognitive, psychological, and even behavioral/physical measurements. They discussed the classic Solomon four-group design as the original method intended to evaluate the effects of a pretest on posttest performance. In this design, four groups are randomized into treatment or control, and within each subgroup, a pretest was given and withheld. Analysis of variance provides results for the effect of the pretest as well as of pretest-treatment interaction. The latter is particularly of interest if it is significant. Depending on the direction of the effect, its presence will indicate the nature of the pretest effect, either enhancing or suppressing the treatment effect in comparison with control conditions. Willson and Putnam reported that pretest effects on cognitive and psychological functioning were positive, enhancing both on posttest scores in comparison to non-pretested groups. Cognitive measurements evoked greater effects than personality or psychological measurements or reports of behavior, but all can substantially alter the interpretation of experimental results if not considered. Their conclusion was that such effects typically peak at about 2 weeks, and after a month or so diminish to a fraction of a standard deviation, perhaps 0.10–0.20 over longer periods. These values are consistent with the reports of the Educational Testing Service for its various tests for test-takers who retake a test after a year. The mechanism for this was not discussed, and in the intervening 25 years, little research has focused on the actual cognitive processes that increase performance. Indirect evidence, however, from well-established research on memory, activation theory, or other cognitive theory such as dual coding can provide rationales for such effects.

Donna T. Heinsman and William R. Shadish examined the effect of pretests on randomized and nonrandomized studies using meta-analysis as part of a broader study of the differences in the two types of designs. They found that greater pretest differences between experimental and control groups were associated with greater differences in outcomes between randomized and nonrandomized designs, with nonrandomized designs producing results too great in comparison. This suggests that simply adjusting the effect statistically using analysis of covariance will not produce good estimates of treatments in nonrandomized designs.

Pretest effects have been found in many contexts. David A. Kravitz and William K. Balzer argued for such effects in nonrandomized experimental studies on performance appraisal used in business. Shadish, G. E. Matt, A. M. Navarro, and G. Phillips reported pretest biases in nonrandomized treatment groups in psychotherapy studies. In an experimental study of pretest format effects as mediators of the relationship between beliefs about tests and test performance, D. Chan, N. Schmitt, J. M. Sacco, and R. P. DeShon conducted an experiment in which all participants were provided example items of the type of test they were to take (both cognitive and psychological) with ratings of the face validity, fairness, and applicability given prior to administration of the actual tests. The structural equation model they analyzed supported a complete mediation model of the effect of beliefs about tests on posttest performance by pretest ratings of the sample items. This indicates that pretest effects can depend on the initial knowledge or belief structure about the type of testing involved. This also suggests that the work by researchers such as Carol S. Dweck and colleagues on academic self-concept and motivation indicated that they may play an important role in pretest effects, but this has not been investigated or established.

## Proposed Methods to Account for Pretest Sensitization

Although the Solomon four-group design was proposed as a standard method to evaluate pretest sensitization, James H. Bray, Scott E. Maxwell, and George S. Howard argued that such designs or even analysis of covariance designs with pretest as covariate failed to properly account for the sensitization that could occur in self-report studies. They proposed that subjects report ex post facto their pretest condition and that researchers use that as a covariate or as a basis for examining treatment differences. Mirjam Sprangers and

Johan Hoogstraten evaluated the effect of retrospective construction of pretests by subjects on communication skills. They concluded that treatments may interfere with subjects' report of their pretreatment condition compared to control subjects, which they discuss as response shift, an internal change in subjects' standards or orientation. Clara C. Pratt, William M. McGuigan, and Aphra R. Katzev found just the opposite, that retrospective pretests prevented response shifts. Response shift analysis is discussed later as a variant on pretest sensitization.

Tony C. Lam and Priscilla Bengo compared three retrospective methods of measuring elementary grade teachers' self-reported change in mathematics instructional practices: the posttest with retrospective pretest method (reporting current practices and earlier practices), the posttest with perceived change method (reporting current practice and the amount and direction of change), and perceived change method (reporting only the amount and direction of change). Teachers in the posttest with retrospective pretest condition reported the least change, followed by teachers in the posttest with perceived change condition, whereas teachers in the perceived change condition reported the greatest change. Their explanation for the findings focused on differential satisficing (responding with minimally adequate rather than optimal effort) of teachers caused by differences in cognitive demands among the three methods. Greater task difficulty leads to greater satisficing, which causes respondents to resort more to socially desirable responses. This study has not yet been replicated, however.

Whether to assess retrospective pretests or not remains inconclusive, perhaps dependent on the nature of the assessment. Subjects are known to be more or less reliable in self-assessments, depending on the type of assessment. The literature on self-report is voluminous, and much too large to summarize here. One example, however, is that children and adolescents are more reliable reporting internalizing problems (and provide better validity evidence than other observers) compared to their self-reports of externalizing problems, and these responses remain reliable across time. Reconstructive memory may play a greater role in experiential retrospective pretests than stable trait assessment, however. The Lam and Bengo result may, in fact, support such reconstruction as much as cognitive demand. This remains an area to investigate.

## Statistical Considerations

Although randomization into treatments as recommended in the Solomon four-group design is desirable, it does not completely account for potential pretest sensitization. The interaction effect of treatment by pretest condition provides an average effect across pretest values, but it does not evaluate differential effects across pretest values. That is, treatment effect differences might vary linearly (or nonlinearly) from low pretest scores to high pretest scores. P. Dugard and J. Todman commented that some researchers simply examine gain scores from pretests, and they argued that this method will produce statistical problems. Functionally, a gain score is simply a regression in which the regression weight is 1.0 rather than the ordinary *b*-weight from least-squares regression. Obviously, such an estimate is potentially biased and inefficient. The best way to examine these effects is to construct within the pretested groups a pretest-treatment interaction variable. Patricia Cohen, Jacob Cohen, Stephen G. West, and Leona S. Aiken discussed constructing interval-category interaction variables by centering the interval variable, constructing a contrast for the treatment variable, and multiplying the two together to produce a new variable that can be included in an analysis of covariance. Because the pretest is available for only half the Solomon design (at least for a balanced, equal sample size per group design), power is reduced, so if one is contemplating this analysis, appropriate adjustments to the sample size of the pretested groups is in order. The design can be modified so that it is proportional, but with a larger sample size in the pretested groups than the unpretested groups. That, in turn, may increase study costs if collecting the pretest data has significant costs associated with it. Still, this approach will permit a more refined examination of the pretest effects on individuals.

Expanding the covariate approach mentioned above, several additional design considerations should be evaluated. First, pretest effects may be nonlinear, such as a quadratic effect that either levels off for high pretest values (possibly because

of ceiling effects of the test) or increases and then decreases with pretest values as treatment becomes less effective for either low or high pretest values. Relatedly, a control group may have a linear relationship with the pretest values whereas the treatment exhibits a quadratic effect. These effects can be modeled.

Yet more elaborate modeling may be important in observational studies using path analysis or structural equation modeling (SEM). Chan et al. used SEM to examine pretesting as a mediator between beliefs and test performance, for example. Pretest score is often included in SEM models as an exogenous predictor. Less frequently, pretest-covariate interaction variables are considered and evaluated, although the potential for their existence and explanatory unpacking of findings is great.

Another approach to pretest sensitization in design of studies is by conducting multiple measurements over time, if it is appropriate. With growth modeling or even time-series modeling of such data, both individuals' and groups' data can be examined for differential intercepts and slopes across the measurements. In this case, the pretest score is ignored as a measurement in the growth or time series design, but pretested and non-pretested groups are analyzed separately, perhaps first as a multiple-group growth or time series model constrained for all parameters equal, then as unconstrained separate models with parameters released as indicated by tests such as chi-square modification indexes in SEM. Of course, such models can be constructed even for single time-point measurement after pretesting, but the capability to examine cross-time differences might provide greater understanding of the pretest effects. Response shifts for internal standards may be examined as a change in intercept between any one time point and the successor; the difficulty is determining when the shift occurred. Using an information measure such as the Akaike information criterion may be helpful in that situation.

## Future Research

Pretest sensitization exists as a predictable effect in social science research design, varying in its strength dependent on the nature of the measurement of the pretest. Although the effect appears to wax and wane over a period of weeks and months, it can produce a long-term albeit small change upward in means of groups receiving pretests. Retrospective reporting provides an alternative that has potential to reduce or remove the pretest effect, but much research remains to determine its salience with respect to type of pretest, length of time required to construct the retrospective estimate, and form of the measurement.

*Victor L. Willson and Eun Sook Kim*

***See also*** Analysis of Covariance (ANCOVA); Experimental Design; Gain Scores, Analysis of; Growth Curve; Interaction; Repeated Measures Design

### Further Readings

Bracht, G., & Glass, G. V (1968). The external validity of experiments. *American Educational Research Journal, 5,* 437–474.

Bray, J. H., Maxwell, S. E., & Howard, G. S. (1984). Methods of analysis with response-shift bias. *Educational and Psychological Measurement, 44*(4), 781–804.

Chan, D., Schmitt, N., Sacco, J. M., & DeShon, R. P. (1998). Understanding pretest and posttest reactions to cognitive ability and personality tests. *Journal of Applied Psychology, 83*(3), 471–485.

Heinsman, D. T., & Shadish, W. R. (1996). Assignment methods in experimentation: When do nonrandomized experiments approximate answers from randomized experiments? *Psychological Methods, 1*(2), 154–169.

Kravitz, D. A., & Balzer, W. K. (1992). Context effects in performance appraisal: A methodological critique and empirical study. *Journal of Applied Psychology, 77*(1), 24–31.

Lam, T. C. M., & Bengo, P. (2003). A comparison of three retrospective self-reporting methods of measuring change in instructional practice. *American Journal of Evaluation, 24*(1), 65–80.

Pratt, C. C., McGuigan, W. M., & Katzev, A. R. (2000). Measuring program outcomes: Using retrospective methodology. *American Journal of Evaluation, 21*(3), 341–349.

Sprangers, M., & Hoogstraten, J. (1989). Pretesting effects in retrospective pretest-posttest designs. *Journal of Applied Psychology, 74*(2), 265–272.

Willson, V. L., & Putnam, R. R. (1982). A meta-analysis of pretest sensitization effects in experimental design. *American Educational Research Journal, 19*(2), 249–258.

# Primary Data Source

A primary data source is an original data source, that is, one in which the data are collected firsthand by the researcher for a specific research purpose or project. Primary data can be collected in a number of ways. However, the most common techniques are self-administered surveys, interviews, field observation, and experiments. Primary data collection is quite expensive and time consuming compared to secondary data collection. Notwithstanding, primary data collection may be the only suitable method for some types of research.

## Primary Data Sources Versus Secondary Data Sources

In the conduct of research, researchers rely on two kinds of data sources—primary and secondary. The term *primary source* is used broadly to embody all sources that are original. This can be contrasted with the term *primary data source,* which refers specifically to the firsthand collection of data for a particular purpose. In the context of the broader definition, primary sources may be published (e.g., census data) or unpublished (e.g., President Lincoln's personal diary), and could be from the past (e.g., artifacts) or present (e.g., poll for a national election). In many types of research, these sources are often created at the time of the event for a specific purpose (e.g., President Obama's inauguration speech, a company's marketing survey of 2009 teenage fashions). However, primary sources may also encompass records of events as described by an eyewitness to the event or someone who actually experienced the event. Both primary data sources and primary sources provide firsthand, unmediated information that is closest to the object of study. This is not a guarantee, however, that these sources are always accurate. For example, if data are collected via face-to-face interviews, the data quality may be affected by a number of issues, such as interviewee characteristics, data transcription, data entry, and so on. Likewise, when eyewitness accounts are given, they can be deliberately or unconsciously distorted.

The delineation of primary and secondary sources first arose in the field of historiography, when historians attempted to classify sources of handwriting. Over time, various disciplines have attempted to normalize the definition for *primary* and *secondary* sources. However, these two classifications remain highly subjective and relative. This is because these terms are highly contextual. A particular source may be considered as either primary or secondary depending on the context in which it is examined.

Primary data sources are most often created using survey research. There are a number of different survey techniques that can be used to collect primary data, such as interviews (e.g., face-to-face, telephone, e-mail, fax) or self-administered questionnaires. When polls, censuses, and other direct data collection are undertaken, these all constitute primary data sources. However, when these sources are subsequently used by others for other research purposes, they are referred to as secondary data sources. Primary data sources may also be created using other methods, such as field observation and experiments (e.g., pretests and test marketing). The latter technique is particularly important in marketing research projects and can be done in either a laboratory or a field setting. In contrast, some common examples of primary sources include speeches, letters, diaries, autobiographies, interviews, official reports, legislation, court records, tax records, birth records, wills, newsreels, artifacts, poetry, drama, films, music, visual art, paintings, photographs, and drawings, along with all those sources that are classified as primary data sources.

Comparatively, a secondary data source refers to a data source that is already in existence (e.g., Gallup, Harris, and Roper polls; General Social Survey; national census) and is being used either for a purpose for which it was not originally intended and/or by someone other than the researcher who collected the original data. Such sources are considered to be nonoriginal or secondhand, but are nonetheless considered to be quite valuable. They are generally obtained from public archives (e.g., Inter-University Consortium for Political and Social Research maintained at the University of Michigan, National Network of State Polls maintained at the University of North Carolina), but may also be obtained from other

researchers or even from a researcher's previous work. Public archives can be accessed and used by researchers either free of cost or for a small fee. The significant cost savings that can result from use of secondary data sources is perhaps one of the greatest attractions of using secondary data. Other advantages include greater efficiency (easier to obtain, less time required) and strengthened confidence (triangulation of primary and secondary sources or multiple secondary sources). However, the challenges associated with secondary data sources, which are discussed later, may far outweigh these benefits. Notwithstanding, these sources are invaluable in certain types of research (e.g., historical and comparative research), and their use, particularly in the United States, has increased substantially over the past 25 years.

## The Need for Primary Data

Researchers collect data for numerous reasons. It may be to answer a particular research question, solve a particular problem, test some hypothesis, validate or falsify an existing theory, better understand some phenomenon, write a report or research paper, or write a thesis or dissertation, or perhaps simply for the sake of learning/knowledge. Regardless of the reason, data collection generally requires a systematic and purposeful approach that is largely determined by the research design best suited to achieve the research objectives. In other words, the research design specifies the type of data that is to be collected, the population and other sources from which the data will be collected, and the procedures that will be used to collect the data.

There are several reasons why researchers may choose to use a primary rather than a secondary data source when conducting research.

*Primary data sources undoubtedly constitute the purest form of data—these sources are firsthand and unfiltered.* As such, primary data sources may be the preferred choice in many types of research, particularly those pertaining to scholarship, where credibility might be questioned if only secondary data sources were used. Additionally, in some types of research such as historical writing, many historians believe that objective connection to the past cannot be captured with secondary sources

*Aggregated data are often not sufficiently detailed to serve the needs of certain types of research, and differences in units of measurement may further complicate the problem.* For example, if a researcher is conducting research at a rural community level, but is using data collected at the country level (e.g., census data), it would be difficult and/or impossible to draw any conclusions at the rural community level unless rural community data could be extracted accurately and the researcher could actually get access to these data. If this is not possible, the researcher would need to gather primary data at the community level to conduct the particular study. Likewise, if a marketing study is being targeted to a particular trade area, but data are aggregated at the town or country level, these data would not be suitable and primary data would need to be collected.

*Regardless of whether researchers can get access to more detailed data, they still need to be concerned about data quality.* The greatest challenge associated with a secondary data source pertains to uncertainty about data collection processes. Even when data are collected by government agencies, this is no indication that the data are of high quality, accurate, or uncontaminated. In order for research to have validity and credibility, the entire research process must be sufficiently rigorous. When researchers collect primary data, they are in complete control of the data collection process. This concern is therefore considerably minimized.

*Researchers may be concerned about the extent to which secondary data sources fit the information needs of the current study, and the related concern of built-in bias.* Data that are collected for one purpose are often not suitable for another. This is not a concern with primary data because data collection is tailored to the specific research objectives, so intended-use bias is eliminated.

*The definition of classes (e.g., age, income) may also present a problem.* For example, in the United States, adults are classified by public opinion polls as persons 18 years or older. However, the specific research study may be interested in adults 21 years or older.

*The age of the data may also be of concern.* Research generally needs current data. However,

due to cost, many national data studies are conducted only every 5 years.

*No secondary data source may exist for some types of research*. For example, scant data have been collected on some ethnic groups (e.g., Hispanics). Thus, if a marketing strategy is being targeted at Hispanics, primary data may be necessary because the spending patterns of Hispanics may be quite different from those of other groups.

*Even if secondary data sources exist that are current, accurate, and relevant, they still may not be suitable for the particular research*. For example, a secondary data source on customer appliance purchase habits would not be suitable for new product development.

*Possible data distortion, which can occur when data are reused multiple times and modified by new authors, may be a concern*. With primary data, distortion is not an issue because the researcher is using original data.

Thus, although secondary data sources are relatively quick and quite inexpensive to obtain, their inherent challenges need to be carefully considered. If these challenges cannot be adequately addressed, the research may produce misleading and inaccurate results. Depending on the type of research that is being conducted, researchers therefore need to answer one or more of the following questions: What were the credentials of the primary investigator? When were the data collected? For whom were the data collected? How were the variables selected? What were they supposed to measure? How was the sample selected? Were the data collection and data entry processes systematic and rigorous? Is the original data collection form available for viewing? Is there someone who can be contacted to provide more insight into the whole data collection and research process? Can less aggregated data be obtained, and what are the ethical implications?

## Ethical Issues Associated With Primary Data Collection

Most types of primary data collection require the approval of an institutional review board (IRB) to ensure the adequate protection of human subjects. Many government departments, universities, and other large institutions generally have their own IRBs. Notwithstanding, the time frame for approval may still take several weeks, especially if a researcher is asked to make changes to his or her application or instrument in order to obtain approval. When secondary data sources are used, researchers are spared the headache of gaining IRB approval. However, other ethical implications may need to be considered if researchers are given access to detailed original data. For example, should participant approval be sought when data that are collected for one purpose are being used for another?

## Cost Implications of Primary Data Sources

Primary data collection is a resource-intense activity in terms of both time and cost. Depending on the purpose for which the data are required, primary data collection may take several months—or years—to accumulate. In addition, the data collection technique chosen can have significant cost implications. For example, face-to-face interviews are considerably more expensive compared to telephone interviews or self-administered surveys because face-to-face interviews not only are longer (approximately 1 hour), but also require the use of more highly skilled interviewers who possess certain characteristics. Other significant costs associated with primary data collection include the cost of data entry and validation and the cost associated with sample selection. In contrast, when secondary data sources are used, the only cost that is generally incurred is the cost to access a database.

*Nadini Persaud*

***See also*** Field Study; Interviewing; Levels of Measurement; Protocol; Secondary Data Source

### Further Readings

Babbie, E. (2004). *The practice of social research* (10th ed.). Belmont, CA: Wadsworth/Thomson Learning.

Steinberg, S. J., & Steinberg, S. L. (2006). *Geographical information systems for the social sciences: Investigating space and place*. Thousand Oaks, CA: Sage.

Weisberg, H. F., Krosnick, J. A., & Bowen, B. D. (1996). *An introduction to survey, research, polling and data analysis* (3rd ed.). Thousand Oaks, CA: Sage.

# Principal Components Analysis

Also known as empirical orthogonal function analysis, principal components analysis (PCA) is a multivariate data analysis technique that is employed to reduce the dimensionality of large data sets and simplify the representation of the data field under consideration. PCA is used to understand the interdependencies among variables and trim down the redundant (or significantly correlated) variables that are measuring the same construct. Data sets with a considerable proportion of interrelated variables are transformed into a set of new hypothetical variables known as principal components, which are uncorrelated or orthogonal to one another. These new variables are ordered so that the first few components retain most of the variation present in the original data matrix. The components reflect both common and unique variance of the variables (as opposed to common factor analysis that excludes unique variance), with the last few components identifying directions in which there is negligible variation or a near linear relationship with the original variables. Thus, PCA reduces the number of variables under examination and allows one to detect and recognize groups of interrelated variables. Frequently, PCA does not generate the final product and is often used in combination with other statistical techniques (e.g., cluster analysis) to uncover, model, and explain the leading multivariate relationships. The method was first introduced in 1901 by Karl Pearson and subsequently modified three decades later by Harold Hotelling for the objective of exploring correlation structures; it has since been used extensively in both the physical and social sciences.

## Mathematical Origins and Matrix Constructs

PCA describes the variation in a set of multivariate data in terms of a new assemblage of variables that are uncorrelated to one another. Mathematically, the statistical method can be described briefly as a linear transformation from the original variables, $x_1, \ldots, x_p$, to new variables, $y_1, \ldots, y_p$ (as described succinctly by Geoff Der and Brian Everitt), where

$$
\begin{aligned}
y_1 &= a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p \\
y_2 &= a_{21}x_1 + a_{22}x_2 + \cdots + a_{2p}x_p \\
&\vdots \\
y_p &= a_{p1}x_1 + a_{p2}x_2 + \cdots + a_{pp}x_p.
\end{aligned}
$$

The coefficients, $a_{pp}$, defining each new variable are selected in such a way that the $y$ variables or principal components are orthogonal, meaning that the coordinate axes are rotated such that the axes are still at right angles to each other while maximizing the variance. Each component is arranged according to decreasing order of variance accounted for in the original data matrix. The number of possible principal components is equal to the number of input variables, but not all components will be retained in the analysis seeing that a primary goal of PCA is simplification of the data matrix (see the subsequent section on principal component truncation methods).

The original coordinates of the $i$th data point, $x_{ij}, j = 1, \ldots, p$, becomes in the new system (as explained by Trevor Bailey and Anthony Gatrell):

$$y_{ij} = a_{j1}x_{i1} + a_{j2}x_{i2} + \cdots + a_{jp}x_{ip}$$

The $j$th new variable $y_j$ is normally referred to as the $j$th principal component, whereas $y_{ij}$ is termed the score of the $i$th observation on the $j$th principal component. The relationship between the $j$th principal component and the $k$th original variable is described by the covariance between them, given as

$$a_{jk}\sqrt{\frac{\lambda_j}{s_{kk}}},$$

where $s_{kk}$ is the estimated variance of the $k$th original variable or the $k$th diagonal element of the data matrix **S**. This relationship is referred to as a loading of the $k$th original variable of the $j$th principal component. Component loadings are essentially correlations between the variables and the component and are interpreted similarly to product-moment correlation coefficients (or Pearson's $r$). Values of components loadings range from –1.0 to 1.0. More positive (negative) component loadings indicate a stronger linkage of a variable on a particular component, and those values closer

to zero signify that the variable is not being represented by that component.

In matrix algebra form, the set of uncorrelated variables is derived by a set of linear transformations, **E**, where

$$\mathbf{\Lambda} = \mathbf{E}^T(Z^T Z)\mathbf{E}.$$

Matrix **E** is the eigenvector matrix extracted from the observed correlation matrix, **R**, and **Λ** is the diagonal matrix of the eigenvalues, $\lambda_i$. The eigenvalue for component *i*, $\lambda_i$, is computed as

$$\lambda_i = \sum_{j=1}^{n} L_{ij}^2,$$

where $L_{ij}$ is the loading for variable *j* on component *i*. It is the sum of the squared loadings that indicates the total variance accounted for by the component in each of the new direction components axes. Each squared component loading signifies the extent to which the new variable or component number is able to characterize the attributes of the original variable. Higher (lower) eigenvalues indicate a larger (smaller) proportion of the total variance. Generally, eigenvalues greater than 1.0 will collectively provide a fitting and efficient description of the data, and thereby reduce the original variable set to a few new hypothetical variables, the principal components.

PCA operates in one of several analytic modes (as outlined by G. David Garson) depending on the data structure or rather the relative positioning of the cases and variables in the data matrix. The most widespread is R-mode, which identifies clusters of variables (the columns) based on a set of cases (the rows) for a given moment in time. R-mode is often the default setting in many statistical packages (e.g., SPSS, an IBM company) and not usually specifically identified as such. In contrast, Q-mode clusters the cases rather than variables and accordingly reverses the row and column inputs of the R-mode. The emphasis in Q-mode is on examining how the cases cluster on particular components and determining if a separation exists among cases based on a positive or negative loading response. Consequently, Q-mode analysis is frequently regarded as a clustering analysis technique and is sometimes used in place of traditional cluster analysis methods (e.g., agglomerative hierarchical cluster analysis).

Other, less common PCA analytic modes include S-mode, T-mode, and O-mode. These modes are usually reserved for PCA involving an extended time period (e.g., several weeks) rather than a single, indeterminate measurement time. In an S-mode data structure, the time component comprises the cases (the rows), the specific data points are the columns, and the individual cells contain information on a single variable. The components produced from an S-mode analysis would demonstrate how these data points cluster together on a variable over time. Spatial data typically have an S-mode construct, where geographically oriented variables (e.g., the daily maximum temperature per station) are examined for similar regional modes of variability across the record period. Alternatively, O-mode PCA analysis looks into time periods that show data points exhibiting similar variances on a particular set of measures. The O-mode PCA structure functions similar to a time series analysis by examining how the time element (the columns) clusters on each component according to a variable set (the rows) for a single case. Like O-mode PCA analysis, the T-mode construct also places emphasis on clustering time on each of the components (the columns), except that information is gathered on a single variable for multiple cases (the rows). The T-mode PCA structure is particularly useful for investigating whether or not differences for a particular variable exist between two or more time periods.

As with all statistical techniques, PCA has a set of requirements and assumptions to be met prior to application. The data set under examination should contain no outliers, lack or diminish selection bias, and assume interval- or ratio-level data with multivariate normality. Furthermore, the underlying dimensions of the input variables are shared, identified as having moderately high intercorrelations that can be determined using Kaiser-Meyer-Olkin (KMO) test statistics. The KMO test statistic of sampling adequacy, based on correlation and partial correlation of the variables, predicts the likelihood that PCA is a viable data reduction technique and will produce meaningful components. The KMO ranges from 0 to 1.0 and should be about 0.60 or higher to proceed with the PCA. In addition to the KMO, Bartlett's Test

of Sphericity (BTS) may also be used to determine if PCA is an appropriate statistical method. BTS tests the hypothesis that the data correlation matrix is an identity matrix, indicating the variables under consideration are unrelated and, hence, inappropriate for structure detection. A BTS test result with a significance level value of 0.05 or smaller indicates that PCA may be a useful data analysis method.

## Principal Component Truncation Methods

Mathematically, the number of eigenvectors produced from the analysis equals the number of variables entered into the data matrix. However, because a major goal of PCA is to simplify the data set while capturing the majority of its variance, the eigenvectors analyzed are often truncated. Typically, the first few principal components capture the most variance in the data set and are retained for further analysis, whereas the other components represent a small, insignificant fraction of the variability. The cumulative percentage of variation is often used to truncate the components after a set percentage threshold has been achieved; components accounting for between 70% and 90% of the total variation in the data set are retained. Several other criteria, which are often rather subjective, have also been used to determine the number of factors to retain from a PCA throughout the physical and social sciences.

One truncation method involves comparing each eigenvalue of each eigenvector to the amount of joint variance shown in the average eigenvalue. The eigenvectors retained are based on a predetermined eigenvalue threshold parameter (e.g., the lowest significant eigenvalue deemed to be meaningful). Commonly, Kaiser's rule is applied, which states that all components with eigenvalues less than 1.0 should be eliminated from the analysis. The rationale behind the Kaiser criterion is that eigenvalues less than 1.0 account for less of the total variance than did any one of the original variables; however, in some cases, it is desirable to select a minimum value lower than 1.0 to allow for sampling variation and a more liberal eigenvalue threshold. For instance, I. T. Jolliffe suggests removing all components with eigenvalues less than 0.7, which can substantially increase the number of retained components and diminish the data reduction capability of PCA.

Graphically based selection methods have the researcher truncate the number of components based on a qualitative assessment of an eigenvalue spectrum plot (i.e., displaying the eigenvalues in descending order). Two such methods are the Cattell scree test (or simply scree plot) and the log-eigenvalue diagram. The scree plot charts the component number ($x$-axis) against the eigenvalues ($y$-axis). Likewise, the log-eigenvalue diagram shows the eigenvalue magnitude as a function of the principal component number, but with a logarithmic transformation to emphasize an exponential drop in the size of the eigenvalues. In both charts, the researcher subjectively examines the plot for a steep decline in the eigenvalues from which there is a flattening of the curve; components immediately prior to the straightening of the curve are retained. The absence of a significant decline or elbow shape to the curve indicates that more than the first few components are needed to adequately represent the variance in the data set.

Communality is also a good indicator of the reliability of the model and the appropriate number of components to keep. It is the squared multiple correlation for a variable using the components as predictors so that

$$h_j^2 = \sum_{i=1}^{k} L_{ij}^2,$$

where $L_{ij}$ is the loading for variable $j$ on component $i$, $k$ is the number of components ($\leq n$), and $h_j^2$ is the communality for variable $j$. The resulting value specifies the proportion of variance accounted for by a particular set of components. Communalities generally range from 0 to 1, where low values indicate that either the variable is not functioning well in the PCA model or additional components need to be extracted in order to capture the variability for that element. Hence, communalities can indicate if an initial truncation was too severe and a more liberal selection criterion should be applied.

In practice, no single truncation method works for all situations, and often, several trials are needed before the final selection of the number of factors to retain is made. For instance,

Kaiser's rule typically retains far too many components and does not adequately simplify the data matrix, especially with an increasing number of input variables. On the other hand, the scree test may retain few components, especially if the first eigenvalue is heavily loaded on the first one or two components. Truncation methods will usually produce the most comparable results when the number of data elements is relatively small in comparison to the number of cases. Moreover, determining the number of components to use is not solely based on obtaining a large proportion of the variance and having each retained component be a significant contributor. The researcher must decide if the retained components are interpretable or meaningful while providing a parsimonious description of the data.

## Eigenvector Rotation

By definition, the eigenvectors produced from PCA are orthogonal to one another, resulting in uncorrelated principal components. This restraint can make interpretation of the principal components difficult, particularly for fields involving fundamental processes that are not mutually exclusive. For example, data sets with a spatial component (e.g., geographical coordinates) often contain elements that are interrelated to one another by virtue of position, such as the temperature of one location being heavily influenced by the temperature, moisture, wind direction, terrain, and so on, of adjacent locations. Although the first component may contain a significant proportion of the variability of a data set, a single process may not actually be represented by the first component and may contain information representing several separate processes that are then difficult to partition out. Hence, unrotated solutions tend to have variables that load on multiple components. The purpose of rotation is to simplify the principal component coefficients and facilitate understanding the primary modes of variability. Rotating the eigenvectors is mainly appropriate when physical interpretation rather than data matrix reduction is the most important result from the PCA.

In nearly all computer statistical packages (e.g., SPSS), the default for PCA is an unrotated solution that will be produced in addition to any selected rotation. Unrotated components will show the maximum variance on the first component, with eigenvalues and associated variance decreasing with each increasing component. In a rotated PCA, the cumulative variance (and same total eigenvalue) produced by the retained principal components remains the same as in the unrotated solution, but the loadings across the individual components change. Accordingly, the first rotated principal component will most likely not represent the maximum variance in the data set.

Rotation methods fall into two main types, orthogonal and oblique, depending on the relative orientation of the eigenvectors. In orthogonal rotation, the eigenvectors are at right angles to one another and remain uncorrelated as in the unrotated PCA. Varimax rotation is the most widespread orthogonal rotation method and seeks to maximize the variance of the squared loadings of a component on all the variables in the component matrix. In other words, the number of variables loading on one component is minimized. On the contrary, oblique rotation techniques produce eigenvectors that are no longer at right angles to one another, thus removing the uncorrelated property between eigenvectors. Direct oblimin and Promax are both oblique methods that usually will simplify the component solution and produce higher eigenvalues in comparison to orthogonal methods; however, interpretability is somewhat diminished.

The decision to rotate the components is often highly dependent on the data structure and individual preference. The usage of the rotated and nonrotated principal components is equally prevalent in the literature and reflects the division within the scientific community. Rotated principal components often are more likely to produce physically meaningful patterns from the data and are more statistically stable, being less sensitive to sampling distribution. Yet as I. T. Jolliffe points out, applying rotation on the principal components has several negative aspects, such as the subjective selection of rotation criterion and method or the sensitivity of the result to the number of original components retained. In particular, a large drawback of rotated PCA

often cited is that the interpretation of the components is highly dependent on the rotation method itself, with each rotation method resulting in new possible explanations. Furthermore, rotation also redistributes the eigenvalues, such that the first few components will no longer consecutively account for the maximum possible variation. Rotated components are usually preferred when difficulty arises in assigning physical meaning to the original retained components, particularly in cases in which the PCA is the terminus of the analysis. Generally, if a large difference in the individual component variance totals is observed between the initial PCA and the rotated PCA, then the rotated component matrix will probably have a more straightforward interpretation than the unrotated matrix.

*Jill S. M. Coleman*

***See also*** Correlation; Covariate; Matrix Algebra; Pearson Product-Moment Correlation Coefficient; SPSS; Variable; Variance

**Further Readings**

Bailey, T. C., & Gatrell, A. C. (1995). *Interactive spatial data analysis*. Essex, UK: Longman.

Der, G., & Everitt, B. S. (2002). *A handbook of statistical analyses using SAS* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Dunteman, G. H. (1989). *Principal components analysis*. Newbury Park, CA: Sage.

Garson, G. D. (2008, March). Factor analysis. *Statnotes: Topics in multivariate analysis* [Online]. Retrieved July 5, 2008, from http://www2.chass.ncsu.edu/garson/pa765/statnote.htm

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology, 24,* 417–441.

Jackson, J. E. (2003). *A user's guide to principal components*. New York: Wiley.

Jolliffe, I. T. (2002). *Principal component analysis* (2nd ed.). New York: Springer.

Manly, B. F. J. (2004). *Multivariate statistical methods: A primer* (3rd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Morrison, D. F. (2005). *Multivariate statistical methods* (4th ed.). Belmont, CA: Brooks/Cole.

Timm, N. H. (2002). *Applied multivariate analysis*. New York: Springer.

Wilks, D. S. (2006). *Statistical methods in the atmospheric sciences* (2nd ed.). London: Elsevier.

# Probabilistic Models for Some Intelligence and Attainment Tests

Often in social science research, data are collected from a number of examinees that are each scored on their performance on items of a test. Here, a "test" broadly refers to an assessment of an examinee's level of ability in a particular domain such as math or reading, or a survey of an examinee's behaviors or attitudes toward something. In *Probabilistic Models for Some Intelligence and Attainment Tests,* George Rasch proposed a model for analyzing such test data. The model, known as the Rasch model, and its extensions are perhaps among the most known models for the analysis of test data. In the analysis of data with a Rasch model, the aim is to measure each examinee's level of a latent trait (e.g., math ability, attitude toward capital punishment) that underlies his or her scores on items of a test. This entry takes a closer look at the Rasch model proposed in *Probabilistic Models for Some Intelligence and Attainment Tests* and its extensions.

Suppose that items of a test are each scored on two levels, say 0 = Incorrect and 1 = Correct, or 0 = False and 1 = True. The Rasch model can be represented by

$$\Pr[X_{ij} = 1|\theta_i, \delta_j] = G(\theta_i - \delta_j), \qquad (1)$$

where $X_{ij}$ is a random variable for the score of examinee $i (i = 1, \ldots, n)$ on the $j$th item of the test $(j = 1, \ldots, J)$, $\theta_i$ is the latent trait parameter of the $i$th examinee, and $\delta_j$ is the difficulty parameter of the $j$th test item. Also, $G(\eta)$ is a cumulative distribution function (c.d.f.), usually assumed to be the c.d.f. of the standard logistic distribution defined by $G(\eta) = \frac{\exp(\eta)}{1+\exp(\eta)}$, although another popular choice, $G$, is the standard normal c.d.f., with $G(\eta)$ = Normal$(\eta|0,1)$. A Rasch model that assumes $G$ to be a standard logistic distribution or a standard normal distribution is called a logistic Rasch model or a normal-ogive Rasch model, respectively.

Many tests contain items that are scored on more than two levels. Suppose that the $j$th item of

a test has possible scores denoted by $k = 0, 1, \ldots, m_j$ for all test items indexed by $j = 1, \ldots, J$. The Rasch partial credit model provides an approach to analyze test items scored on multiple levels, given by

$$\Pr[X_{ij} = k|\theta_i, \delta_{jk}, k = 1, \ldots, m_j] = \frac{\prod_{h=0}^{k} G(\theta_i - \delta_{jh}) \prod_{l=k+1}^{m_j} [1 - G(\theta_i - \delta_{jl})]}{\sum_{y=0}^{m_j} \prod_{h=0}^{y} G(\theta_i - \delta_{jh}) \prod_{l=y+1}^{m_j} [1 - G(\theta_i - \delta_{jl})]}, \quad (2)$$

for examinees $i = 1, \ldots, n$ and test items indexed by $j = 1, \ldots, J$, where $\delta_{jk}$ represents the difficulty of attaining the $k$th score level in item $j$ [fixing $G(\theta_i - \delta_{j0}) \equiv 1$, for all $j = 1, \ldots, J$]. The Rasch rating scale model is a special case of the partial credit model, where it is assumed that all items of a test are scored on the same number of levels, with the levels having difficulties that are the same across items, that is, $m_j = m\ (>1)$ and $\delta_{jk} = \tau_k$ for all test items $j = 1, \ldots, J$ and all score levels $k = 0, 1, \ldots, m$. Note also that the Rasch model for two-level items, presented in Equation 1, is also a special case of the partial credit model, where $m_j = 1$ and $\delta_{jk} = \delta_j$ for all test items $j = 1, \ldots, J$. Henceforth, to maintain simplicity in discussion, the item parameters of a Rasch model are represented by $\delta_{\mathrm{jk}}$ (for all $k = 1, \ldots, m_j$ and all $j = 1, \ldots, J$), with the understanding that $\delta_{jk}$ has a specific form for either the dichotomous, partial credit model, or rating scale Rasch model, as mentioned.

There are many ways to extend and increase the flexibility of the Rasch model, and a few of them are mentioned here. For example, the data analyst can incorporate predictor variables (e.g., the gender of the examinee, or the time that the item was administered, or the judge who rated the examinee), and specify $G(\theta_i - \delta_{jk} + \beta_1 x_{1i} + \cdots + \beta_p x_{pi})$ in the model. Also, the multidimensional random coefficients multinomial logit model provides a general class of extended Rasch models that can be used to parameterize multidimensional examinee latent traits or multidimensional items. Furthermore, it is assumed that each examinee falls into one of a number of latent classes in a population indexed by $c = 1, \ldots, C$, so it is possible to extend to a latent-class Rasch model by taking $G(\theta_{ic} - \delta_{jkc})$, with $\theta_{ic}$ the examinee latent trait and $\delta_{jkc}$ the item parameter in the $c$th latent class. Alternatively, using ideas of Bayesian inference, it is also possible to conduct a latent-class Rasch analysis, without assuming that the number of groups $C$ is known a priori, by taking $G(\theta_{ij} - \delta_{ijc})$, with the distribution of the examinee and/or item parameters modeled nonparametrically with a Dirichlet process prior distribution. Finally, instead of assuming that $G$ is a fixed and known monotonic function (e.g., logistic c.d.f.), as is done with all the Rasch models and extensions mentioned above, a more flexible Rasch model can be defined by modeling the c.d.f. $G$ nonparametrically as an unknown parameter, by specifying a nonparametric prior distribution on $G$, such as the Dirichlet process prior.

Given a sample of test data, there are several ways to estimate the (population) parameters of a Rasch model. Let $\omega$ denote the vector of all parameters in a given Rasch model. Although maximum likelihood estimation can be used to obtain a point-estimate $\hat{\omega}$, it does not yield a consistent estimate because in the Rasch model, the number of examinee parameters grows with the sample size. This problem is addressed by either one of two estimation methods that each eliminate examinee parameters, and find maximum likelihood estimates of the remaining parameters of the model. The first method is marginal maximum likelihood estimation, where the examinee latent trait parameters are eliminated by marginalizing over an assumed normal distribution of latent traits in the population. The second method is conditional maximum likelihood estimation, where the latent trait parameter of each examinee is eliminated by conditioning on his or her total test score (a sufficient statistic in the Rasch model). Finally, through the use of Bayes's theorem, estimation of the Rasch model is carried out through inference of the full posterior distribution of $\omega$, which is based on a prior distribution on $\omega$ that is updated by the observed data.

*George Karabatsos*

***See also*** Bayes's Theorem; Estimation; Item Response Theory; Nonparametric Statistics; Test

**Further Readings**

Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21,* 1–23.

Andersen, E. B. (1970). Asymptotic properties of conditional maximum-likelihood estimators. *Journal of the Royal Statistical Society, Series B, 32,* 283–301.

Andrich, D. (1978). A rating fomulation for ordered response categories. *Psychometrika, 43,* 561–573.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46,* 443–459.

Embretson, S. E., & Riese, S. P. ( 2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum.

Fischer, G. H., & Molenaar, I. W. (1995). *Rasch models: Foundations, recent developments and applications.* New York: Springer-Verlag.

Karabatsos, G., & Walker, S. G. (2009). Coherent psychometric modeling with Bayesian nonparametrics. *British Journal of Mathematical and Statistical Psychology, 62,* 1–20.

Kleinman, K. P., & Ibrahim, J. G. (1998). A semi-parametric Bayesian approach to generalized linear mixed models. *Statistics in Medicine, 17,* 2579–2596.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149–174.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Paedagogike Institut.

Von Davier, M., & Carstensen, C. H. (2006). *Multivariate and mixture distribution Rasch models: Extensions and applications.* New York: Springer.

# Probability, Laws of

The laws of probability are a collection of equations that define probabilistic relationships among events. The validity of each equation, or probability law, often depends on the events having met certain conditions. Thus, each law is accompanied by a list of what, if anything, must be assumed about the events in question for the law to hold. There is no specific number or set of equations defined by the term *laws of probability.* A list that covers probabilistic relationships that are most often of interest, such as the probability that at least one of two events occurs, or the probability that two events occur together, is given here. Probability laws are employed frequently in quantitative research, as they are an essential piece of the foundation upon which much of the field of statistics—and all statistical inference—is built.

## Experiments, Outcome Sets, and Events

Probability laws are defined in the context of an *experiment* whose outcome is due to chance. The set of all possible outcomes for a particular experiment is the *outcome set,* or *sample space,* and is commonly denoted by $\Omega$. An *event* is any subset of the outcome set $\Omega$. Examples of experiments and their corresponding outcome sets are given below; each is accompanied by two event examples.

*Experiment 1:* Flip a coin and record which side, heads or tails, lands face up. In this case, the sample space consists of two elements: heads and tails. The outcome set could be denoted $\Omega = \{H, T\}$. If $E_1$ represents the subset {T} of $\Omega$, then $E_1$ is the event that a tail is flipped. If $E_2$ represents the subset {H} of $\Omega$, then $E_2$ is the event that a head is flipped.

*Experiment 2:* Flip a coin twice and record the side that lands face up in the first flip, and then in the second flip. Here, there are four possible outcomes: $\Omega = \{\text{HH, HT, TH, TT}\}$, where, for example, TH represents the outcome where a tail is observed on the first flip and a head on the second. If $E_1 = \{\text{HH, TT}\}$, then $E_1$ denotes the event that the two flips are in agreement. If $E_2 = \{\text{TH, TT}\}$, then $E_2$ denotes the event that a tail was flipped first.

*Experiment 3:* Flip a coin until the first head appears. In this case, the number of outcomes is infinite, but countable: $\Omega = \{\text{H, TH, TTH, TTTH}, \ldots\}$. If $E_1$ is the event that at most two tails are observed before the first head, then $E_1$ is equal to the subset {H, TH, TTH} of $\Omega$. If $E_2$ is the event that exactly three tails are observed, then $E_2 = \{\text{TTTH}\}$.

*Experiment 4:* Roll a six-sided die so that it comes to rest on a table. Record the number of spots facing up. In this example, the outcome set is numeric: $\Omega = \{1, 2, 3, 4, 5, 6\}$. If $E_1$ is the event that a number greater than 2 is rolled, then

$E_1 = \{3, 4, 5, 6\}$. If $E_2$ is the event that an even number is rolled, then $E_2 = \{2, 4, 6\}$.

*Experiment 5:* Roll a six-sided die so that it comes to rest on a round table. Record the distance from the die to the center of the table. If $r$ is the radius of the table, then the experimental outcome must be a real number from 0 to $r$; hence, the outcome set is infinite and uncountable: $\Omega = [0, r)$. Notice that this interval implies that a die rolling a distance $r$ or greater from the center of the table will fall off the table (and hence be ignored as it does not fit the definition of this experiment). If $E_1$ is the event that the die lands closer to the center of the table than to the nearest edge, then $E_1 = [0, r/2)$. If $E_2$ is the event that the die lands in the outer half of the table area, then $E_2 = \left[(r/\sqrt{2}), r\right)$.

These experiments exemplify a variety of outcome set characteristics: They can be numeric or non-numeric, as well as finite, countably infinite, or uncountably infinite. Probability laws, which govern the probabilities assigned to subsets of $\Omega$, are invariant to such outcome set distinctions.

## Union, Intersection, and Disjoint Events

The *union* of two events $E_1$ and $E_2$, denoted $E_1 \cup E_2$, is the subset of $\Omega$ such that each element of this subset is in $E_1$, $E_2$, or both $E_1$ and $E_2$. Because unions of events are also subsets of $\Omega$, event unions are themselves events. The following four expressions of event union are equivalent:

- $E_1 \cup E_2$
- the union of $E_1$ and $E_2$
- either $E_1$ or $E_2$ occurs
- at least one of $E_1$ or $E_2$ occurs.

Note that "either $E_1$ or $E_2$ occurs" includes cases where both $E_1$ and $E_2$ occur.

The *intersection* of two events $E_1$ and $E_2$, denoted $E_1 \cap E_2$, represents only the elements of $\Omega$ common to both $E_1$ and $E_2$. As with unions, event intersections are subsets of $\Omega$ and therefore are events themselves. The following three expressions of event intersection are equivalent:

- $E_1 \cap E_2$
- the intersection of $E_1$ and $E_2$
- both $E_1$ and $E_2$ occur.

If events $E_1$ and $E_2$ do not intersect, then these two events cannot both be observed in a single experimental outcome. In this case, $E_1$ and $E_2$ are said to be *disjoint,* or *mutually exclusive*.

In the single coin flip example, Experiment 1, the union of events $E_1$ and $E_2$ represent the entire outcome set: $E_1 \cup E_2 = \{H, T\} = \Omega$. The intersection of these two events, however, is empty; these events are disjoint. If $\varnothing$ denotes the empty set, then $E_1 \cap E_2 = \varnothing$.

In the double coin flip example, Experiment 2, the union of events $E_1$ and $E_2$ is a three-element subset of $\Omega$: $E_1 \cup E_2 = \{HH, TH, TT\}$. The intersection of these two events is a single outcome: $E_1 \cap E_2 = \{HT\}$. Because this intersection is not empty, $E_1$ and $E_2$ are not disjoint.

The union of events $E_1$ and $E_2$ in the first head coin flip example, Experiment 3, contains four outcomes: $E_1 \cup E_2 = \{H, TH, TTH, TTTH\}$. Because no elements of $E_1$ are common to $E_2$, the intersection of these two events is empty and the events are disjoint: $E_1 \cap E_2 = \varnothing$.

In the spot-recording die example, Experiment 5, rolling a number greater than 2($E_1$) or rolling an even number ($E_2$) is an event composed of four outcomes: $E_1 \cup E_2 = \{3, 4, 5, 6\}$. These events are not disjoint, as their intersection is not empty: $E_1 \cap E_2 = \{4, 6\}$.

In the distance-recording die example, Experiment 5, rolling closer to the center of the table than to the nearest edge ($E_1$) or rolling in the outer half of the table area ($E_2$) is an event composed of two nonoverlapping intervals: $E_1 \cup E_2 = \{[0, r/2), [(r/\sqrt{2}), r)$. Because there are no elements of $\Omega$ common to both $E_1$ and $E_2$, these events are disjoint: $E_1 \cap E_2 = \varnothing$.

## Three Axioms of Probability

Suppose a particular experiment is performed many, many times. Then the *probability* of an event $E$, denoted $P(E)$, is the proportion of those times that this event would be expected to occur. In other words, the probability of an event is the expected long-run relative frequency of the event. Not only do the laws of probability govern the values $P(E)$ may assume, they also provide the

means to equivalently express probabilities involving multiple events.

The following three probability laws—called *axioms* because they are sufficiently self-evident—provide the basis for proving all other probability laws:

1. $P(E) \geq 0$ for any event $E$.
2. $P(\Omega) = 1$.
3. $P(E_1 \cup E_2) = P(E_1) + P(E_2)$, provided $E_1$ and $E_2$ are disjoint.

The first law states that probabilities can never be negative. This follows directly from the expected long-run relative frequency interpretation of probability. Probabilities can, however, be zero. Events having zero probability can never be observed as an experimental outcome. Events having probability one, on the other hand, are always observed on *every* experimental outcome (this also follows directly from the expected long-run relative frequency interpretation of probability). Because $\Omega$ is composed of all possible experimental outcomes, some element of $\Omega$ must be observed every time the experiment is performed. Hence, $P(\Omega) = 1$. Furthermore, the combination of Axioms 1 and 2 reveal that, for any event $E$, the probability of this event must be contained in the interval [0, 1].

The third axiom, or law, is valid as long as events $E_1$ and $E_2$ are disjoint. Provided they are disjoint, the probability of the union of these two events is equal to the sum of their individual probabilities. This follows from the expected long-run relative frequency probability interpretation: The proportion of times $E_1$ or $E_2$ occurs is equal to the proportion of times $E_1$ occurs, plus the proportion of times $E_2$ occurs, assuming these events cannot happen simultaneously. This axiom also implies that the union of any number of disjoint events has probability equal to the sum of the individual event probabilities.

## Complementary Events and Partitions

For any event $E$, define the *complement* of $E$, denoted $E^C$, to be all elements of $\Omega$ not contained in $E$. Notice $E^C$ is a subset of $\Omega$, and therefore is an event. This implies that both of the following are true:

- $E \cup E^C = \Omega$, and
- $E \cap E^C = \varnothing$.

Any collection of two or more disjoint events (all pairwise intersections are empty) whose union represents the entire outcome set is called a *partition* of $\Omega$. Thus, an event and its complement will always form a partition of $\Omega$. An example of a three-event partition can be found using the spot-recording die example, Experiment 4, and letting $E_3$ be the event a 2 is rolled, and $E_4$ be the event a 1 is rolled. Then, having defined $E_1 = \{3, 4, 5, 6\}$, the events $E_1$, $E_3$, and $E_4$ form a partition of $\Omega$: Any pair of these three events is disjoint, and $E_1 \cup E_3 \cup E_4 = \Omega$.

## Complementary Event Probability

The probability of $E^C$ is found easily by using probability Axioms 2 and 3. By definition of complementary events, $\Omega = E \cup E^C$. The probabilities of these identical events are also identical: $P(\Omega) = P(E \cup E^C)$. Axiom 2 allows us to express $P(\Omega)$ as 1, giving $1 = P(E \cup E^C)$. Axiom 3 allows us to express $P(E \cup E^C)$ as $P(E) + P(E^C)$ because $E$ and $E^C$ are disjoint. This gives $1 = P(E) + P(E^C)$, and thus

$$P(E^C) = 1 - P(E) \text{ for any event } E.$$

Sometimes referred to as the *complementary law of probability,* this rule states that the chance an event does not occur is equivalent to one minus the probability that the event does occur.

## General Probability of Event Union

The general fomula for the probability that $E_1$ or $E_2$ occurs is

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$$
$$\text{for any events } E_1 \text{ and } E_2.$$

This general rule, sometimes referred to as the *additive law of probability,* can be proven using Axiom 3 in conjunction with the complementary law of probability. Notice that Axiom 3 is a special

case of this additive law: If $E_1$ and $E_2$ are disjoint, then their intersection is empty and has probability zero.

### Event Probability in Terms of a Partition

Suppose $E_1, E_2, \ldots, E_k$ represents a $k$-event partition of $\Omega$. Then, for any event $F$, the events $F \cap E_1$ and $F \cap E_2$ are disjoint, as the part of $F$ that overlaps with $E_1$ must have nothing in common with the part of $F$ that overlaps $E_2$ ($E_1$ and $E_2$ are, by definition of a partition, disjoint). Hence any two of the events $F \cap E_1$, $F \cap E_2$, . . . , $F \cap E_k$ are disjoint. Furthermore, the union of these events must equal $F$,

$$F = (F \cap E_1) \cup (F \cap E_2) \cup \ldots \cup (F \cap E_k),$$

because $E_1 \cup E_2 \cup \ldots \cup E_k = \Omega$. This allows for a rule that expresses the probability of an event $F$ in terms of a partition:

$$P(F) = P(F \cap E_1) + P(F \cap E_2) + \cdots + P(F \cap E_k) \quad \text{for any event } F,$$

provided the events $E_1, E_2, \ldots, E_k$ form a partition of $\Omega$. There is no commonly used nomenclature for this probability law.

### Conditional Probability

Conditional probability asks the question: If an experiment is performed and it is only known that event $E$ has occurred, then what is the chance that event $F$ has also occurred? The common notation for a conditional probability is $P(F|E)$, and is read as "the (conditional) probability of $F$, given $E$."

The key to computing conditional probabilities is recognizing that because event $E$ is known to have occurred, the outcome set for that particular experiment can be *reduced* to the subset $E$. If no elements of $F$ are contained within $E$, and $E$ has occurred, then $F$ cannot have occurred. Thus, the conditional probability of $F$, given $E$, is always zero when $F$ and $E$ are disjoint. If the intersection of $F$ and $E$ is not empty, then some elements of $E$ are also elements of $F$. The probability of these common elements, divided by the probability of $E$ itself, gives the *law of conditional probability:*

$$P(F|E) = \frac{P(F \cap E)}{P(E)} \quad \text{for any event } F, \text{ and any event } E \text{ where } P(E) > 0.$$

In the spot-recording die example, Experiment 4, suppose the die is rolled and it is known only that an even number of spots is showing (event $E_2$ is known to have occurred). Then the conditional probability that the roll is a 2 (event $E_3$) is given by

$$P(E_3|E_2) = \frac{P(E_3 \cap E_2)}{P(E_2)} = \frac{P(\{2\} \cap \{2, 4, 6\})}{P(\{2, 4, 6\})} = \frac{P(\{2\})}{P(\{2, 4, 6\})}.$$

Thus, the chance of having rolled a 2, given that the roll was even, is the probability of rolling a 2 divided by the probability of rolling an even number. If we assume that all six faces of the die are equally likely (probability 1/6 each), then this probability is 1/6 divided by 3/6, which equals 1/3. Note that without knowledge that an even number was rolled, the probability of a 2 is 1/6. With knowledge that $E_2$ has occurred, however, the probability of rolling a 2 is updated to 1/3.

Consider the slight change to this example where the roll is still known to have been even, but now the conditional probability of rolling greater than 2 (event $E_1$) is of interest. The conditional law of probability gives

$$P(E_1|E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)} = \frac{P(\{3, 4, 5, 6\} \cap \{2, 4, 6\})}{P(\{2, 4, 6\})} = \frac{P(\{4, 6\})}{P(\{2, 4, 6\})}.$$

Assuming the die to be fair, the probability of rolling a 4 or 6 is 2/6, whereas the probability of rolling an even number is 3/6. Hence, the chance of rolling greater than 2, given that an even number was rolled, is 2/6 times 6/3, which equals 2/3. Note, then, that the chance of rolling greater than 2 without knowledge of rolling an even number, is also 2/3. This example can be expressed symbolically as $P(E_1|E_2) = P(E_1) = 2/3$.

## Independence

In cases where $P(E_1|E_2) = P(E_1)$, the probability that $E_1$ has occurred does not change with knowledge that $E_2$ has occurred. Whenever a pair of events exhibits such a relationship, they are said to be *independent*. To check whether events are independent, it is common to calculate the probabilities on both sides of this equation, or its equivalent counterpart $P(E_2|E_1) = P(E_2)$, and check for equality. If equality does not hold, then the events are said to be dependent.

Another equivalent, and more common, definition of independent events can be found by multiplying both sides of $P(E_1|E_2) = P(E_1)$ by $P(E_2)$. Then use the conditional probability law to recognize the resulting left-hand side as $P(E_1 \cap E_2)$. This gives the following definition of independent events:

> If events $E_1$ and $E_2$ satisfy
> $$P(E_1 \cap E_2) = P(E_1)P(E_2),$$
> then $E_1$ and $E_2$ are independent events.

The converse of this statement also holds, and is sometimes referred to as the *multiplicative law of probability:*

> $P(E_1 \cap E_2) = P(E_1)P(E_2)$ when
> $E_1$ and $E_2$ are independent.

This law is relied upon frequently, not necessarily to check for independence, but rather to calculate the probability of event intersections upon *assuming* independence. Furthermore, if three events are to be independent, then not only must the probability of their three-way intersection equal the product of their three individual probabilities, but the intersection probability of any pair of events must equal the product of the corresponding two individual probabilities.

Note that the term *independent* should not be confused with *disjoint*. The probability of disjoint events occurring simultaneously is zero, whereas the probability of independent events occurring simultaneously is—for all cases of practical relevance—strictly positive.

## Law of Total Probability

The law of total probability uses the conditional probability law to modify the probability of an event $F$ in terms of a partition. If $E_1, E_2, \ldots, E_k$ form a partition of $\Omega$, then $P(F \cap E_1)$ can be equivalently expressed as $P(F|E_1)P(E_1)$ by the conditional probability law. This gives the following equation

$$P(F) = P(F|E_1)P(E_1) + P(F|E_2)P(E_2) + \cdots + P(F|E_k)P(E_k),$$

which is commonly known as the *law of total probability*. This law is useful in situations where the probability of an event of interest ($F$) is known within each partition of the outcome set, but not overall. For example, consider a population of individuals where 20% are considered to be young, 50% are considered to be middle-aged, and the remaining 30% are considered to be senior. Suppose it is known that 10% of those who are young have condition $F$, 30% of those who are middle-aged have condition $F$, and 70% of those who are senior have condition $F$. The probability of selecting an individual at random from this population (put their names in a hat and draw one, for example) who has condition $F$ is given by the law of total probability. Letting $E_1$ be the event that a youth is selected, $E_2$ be the event that a middle-aged individual is selected, and $E_3$ be the event that a senior is selected gives

$$\begin{aligned} P(F) &= P(F|E_1)P(E_1) + P(F|E_2)P(E_2) + P(F|E_3)P(E_3) \\ &= (0.1)(0.2) + (0.3)(0.5) + (0.7)(0.3) \\ &= 0.38. \end{aligned}$$

Thus, there is a 38% chance that an individual selected from this population will have condition $F$. Equivalently, 38 out of every 100 individuals selected (with replacement) from this population are expected to have condition $F$.

## Other Laws of Probability

Any equality or inequality regarding the probability of an event (or events) can be considered a law of probability, provided it can be derived from the three probability axioms. If, for example, event $E_1$

is a subset of event $E_2$, then the relationship $P(E_1) \leq P(E_2)$ must hold. The laws presented previously, however, can be considered among those of most practical use.

Finally, it should be cautioned that events whose probabilities satisfy probability laws are *not* necessarily events whose probabilities are realistic. One may decide, in the spot-recording die example, Experiment 4, that the probabilities associated with rolling numbers 1 through 5 are all 0.1, while the probability associated with rolling a 6 is 0.5. Such an assignment of probabilities (or probability model) is entirely legitimate according to the laws of probability. This assignment would not, however, describe well the outcomes of this experiment if the die were fair (or biased in any way other than what these probabilities represent). Although probability laws can be used to calculate correct event probabilities, it is up to the researcher to ensure that the components used for such calculations are realistic.

*John C. Kern II*

***See also*** Bayes's Theorem; Markov Chains; Random Variable

**Further Readings**

DeGroot, M. H. (1986). *Probability and statistics* (2nd ed.). Reading, MA: Addison-Wesley.
Karr, A. F. (1993). *Probability.* New York: Springer-Verlag.
Kelly, D. G. (1994). *Introduction to probability.* New York: Macmillan.
Ross, S. (2006). *A first course in probability* (7th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.
Ross, S. (2007). *Introduction to probability models* (9th ed.). Burlington, MA: Academic Press.

# Probability Sampling

In many scientific inquiries, it is impossible to individually appraise all the elements that comprise a population of interest. Instead, there is a need to infer collective properties of the whole population from a select subset of its elements. Probability sampling is an approach to selecting elements from a fixed population in such a way that

1. elements are selected by a random process,
2. every element has a nonzero chance of selection, and
3. the relative frequency with which an element is included in a sample is deducible.

A collection of elements drawn in such a way is referred to as a probability sample. The first condition imparts a degree of objectivity to a probability sample and, in combination with the other two conditions, secures a basis for statistical inferences concerning descriptive parameters of the population. It is also largely because its integrity rests on these conditions of the selection process, rather than on the acuity of the investigator, that probability sampling is so widely adopted in modern scientific and statistical surveys.

## Sample Selection

In probability sampling, the selection probabilities of individual population elements and the algorithm with which these are randomly selected are specified by a sampling design. In turn, to apply a sampling design requires a device or frame that delineates the extent of the population of interest. A population often can be sampled directly using a list frame that identifies all the elements in that population, as when the names of all students in a district are registered on school records. If a list frame is available, a probability sample can be formed as each of a string of random numbers generated in accordance with a design is matched to an element or cluster of elements in the population. By contrast, some populations can be framed only by their spatial boundaries. For example, many natural resource populations can be delineated only by the tracts of land over (or under) which they are dispersed. Such a population must be surveyed indirectly via a probability sampling of coordinate locations within its spatial domain. Regardless of how a population is framed, however, the frame must be complete in the sense that it includes the entirety of the population. Inasmuch as any fraction of the population omitted from the frame will have zero probability of being selected, the frame ultimately fixes

the population to which probability sampling inferences apply.

In some applications, once a design is chosen, the set of all selectable probability samples can be discerned together with the relative frequency with which each sample will be drawn. In other cases, the size of the population is never known and one cannot calculate even the number of possible samples. Nevertheless, when a valid probability sampling design is employed, at a minimum it is possible to derive the inclusion probabilities of the elements ultimately selected. The inclusion probability of an element is the relative frequency with which it is included in the observation set. Some probability sampling designs prescribe equal inclusion probabilities for all elements, but most allow for variation across the population. Designs of the latter variety intrinsically favor the selection of certain elements or classes of elements, but only to an extent that can be discerned from the relative magnitudes of the elements' inclusion probabilities.

Notably, even where a design skews the inclusion probabilities toward a certain class of elements, random selection precludes both the investigator (and the elements themselves) from directly influencing the composition of a probability sample. To underscore this point, it is informative to contrast probability sampling with non-probability sampling methods, such as purposive or quota sampling. In these strategies, the investigator plays a direct role in the selection process, often with the aim of assembling a sample that is in some sense representative or typical of the population. By definition, population elements are not drawn objectively, and the likelihood of having observed one element as opposed to another is inscrutable. A non-probability sample can yield an estimate that is close in value to a particular population parameter, but because of its discretional nature, the selection process provides no basis for assessing the potential error of that estimate. In counterpoint, probability sampling is not designed to select representative samples, save in the weak sense of allowing every element a nonzero probability of being observed. Probability sampling is instead fomulated to objectify the selection process so as to permit valid assessments of the distribution of sample-based estimates.

## Estimation and Inference

Applied to a given population, probability sampling admits the selection of any one of a typically large number of possible samples. As a result, a chosen parameter estimator acquires a distribution of possible values, each value corresponding to at least one selectable sample. This distribution is often referred to as the randomization distribution of an estimator because it arises from the random selection process. In practice, an estimator takes only one value when applied to the sample data actually collected, but this singular estimate is nonetheless a realization of the randomization distribution induced by the sample design.

The properties of an estimator's randomization distribution are conditioned by various aspects of the population being sampled but are also controlled by the sampling design. Indeed, an advantage of probability sampling is that the character and degree of variability in an estimator's randomization distribution often can be derived from the sampling design and the estimator's algebraic form. In particular, without setting conditions on the structure of the population, one can generally specify estimators that are unbiased for specific parameters. For example, take the Horvitz–Thompson estimator,

$$\hat{\theta} = \sum_{k \in s} \frac{y_k}{\pi_k},$$

which expands the attribute measurement $y_k$ made on the $k$th distinct element in a probability sample $s$ by that element's inclusion probability ($\pi_k$) and sums these expansions across all selected elements. This estimator is unbiased for the true population total of the $y_k$ under any probability sampling design. Thus, to unbiasedly estimate total research expenditures in a particular sector of the economy, a probability sample of firms in that sector could be drawn and then the sum of the ratios of each firm's research outlay ($y_k$) to inclusion probability ($\pi_k$) could be calculated.

Modifications of the Horvitz–Thompson and other estimation rules provide unbiased estimators of population means, proportions, and distribution functions. Additionally, for many designs, unbiased estimators of the variance of these parameter estimators can be derived. Hence, with many

probability sampling strategies, one can estimate population parameters of interest in an unbiased manner and attach valid measures of precision to those estimates. Going further, one can draw inferences on parameters via confidence intervals that account for the variability of the estimator, given the sampling design.

## Designs

The random selection of elements in probability sampling introduces the possibility of obtaining estimates that deviate markedly from the parameter(s) of interest. Fortunately, there exist numerous design features that dampen the magnitude of sample-to-sample variation in estimates and reduce the likelihood of large deviations. Broadly, the accuracy of an estimator can be improved by applying designs that enhance the dispersion of sample elements across the population or that exploit pertinent auxiliary information on the structure of that population.

Simple random sampling (SRS) is among the most basic of probability sampling designs. From a list frame, a fixed number $n$ of independent selections are made with all elements having identical inclusion probabilities. Whether elements are selected with or without replacement, SRS results in every combination of $n$ elements being drawn with the same frequency. Without replacement, SRS carries the advantage that, on average, individual samples contain a larger number of distinct population elements. Therefore, more information about the population is collected than when element-replacement is permitted, and there is less sample-to-sample variation among parameter estimates.

By eliminating the independence of selections, systematic sampling provides a direct means of drawing dispersed sets of elements. With a systematic probability sampling design, one element is randomly selected from the frame, and then all elements that are separated from this initial selection by a fixed sampling interval are added to the sample. A sample of citizens voting at one location, for example, could be drawn by rolling a die to select one of the first six voters and then subsequently interviewing every sixth voter to exit the polling station. Systematic sampling generally ensures that all elements have equal inclusion probabilities but, at the same time, renders observable only those combinations of elements that are congruent with the sampling interval. Systematic designs are particularly efficient when the sampling interval separates selected elements along a population gradient, be it a natural gradient or one artificially imposed by ordering the frame. By ensuring that every possible sample spans such a gradient, a systematic design reduces variation among estimates and generally improves precision relative to SRS. Unfortunately, because sample elements are not selected independently, it is difficult to assess the precision of estimates if only a single systematic sample is collected.

If prior information exists on natural classifications of the elements of a population, a stratified sampling design can be employed. Such a design divides the population into an exhaustive set of disjoint strata and specifies the collection of mutually independent probability samples from each. Thus, pupils in a school district or trees in a forest stand might be stratified into elementary and secondary students, or into conifer and deciduous species, prior to sampling. As sample elements are necessarily drawn from each stratum, stratified designs permit estimation of stratum-level parameters while ensuring a broad level of coverage across the population as a whole. Moreover, if a significant proportion of the variability in the attribute of interest (e.g., student attendance rate or tree biomass) is due to differences among strata, stratification can lead to large gains in precision relative to SRS.

Stratified sampling designs often vary the inclusion probabilities across strata in order to sample larger, more variable, or more important strata with higher intensity. This concept is carried further by many unequal probability sampling designs, such as Poisson sampling and list sampling. These designs are most effective when the inclusion probability of each element can be made approximately proportional to the magnitude of the attribute of interest. Often this is achieved by making the inclusion probabilities proportional to a readily available auxiliary variable that is in turn positively correlated with the attribute of interest. Thus, if interest centers on total research expenditures in a manufacturing sector, firms with a larger number of employees might be assigned larger inclusion probabilities; if total tree biomass in

a stand is of interest, trees with large basal diameters might be assigned larger inclusion probabilities. The attribute measurements taken on the sampled elements can be weighted by their respective inclusion probabilities to secure unbiased estimation. Additionally, if the inclusion probabilities have been judiciously chosen, the probability-weighted attribute measurements can be appreciably less variable than the raw attribute scores, which can lead to substantially less variation among estimates than is seen under SRS.

One area of active research in probability sampling is the incorporation of statistical models into sample selection and estimation strategies. In many cases, this offers the potential to improve accuracy without sacrificing the objectivity of the probability sampling design as the basis for inference. Of course, in many applications, statistical models can be used to great effect as a basis for inference, but the validity of inferences so drawn then rest on the veracity of the presumed model rather than on the sample selection process itself.

*David L. R. Affleck*

***See also*** Estimation; Nonprobability Sampling; Parameters; Population; Random Sampling

**Further Readings**

Gregoire, T. G., & Valentine, H. T. (2008). *Sampling strategies for natural resources and the environment.* Boca Raton, FL: Chapman & Hall/CRC.

Overton, W. S., & Stehman, S. V. (1995). The Horvitz-Thompson theorem as a unifying perspective for probability sampling: With examples from natural resource sampling. *American Statistician, 49,* 261–268.

Särndal, C.-E., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling.* New York: Springer.

# "Probable Error of a Mean, The"

Initially appreciated by only a handful of brewers and statisticians, "The Probable Error of a Mean" is now, 100 years later, universally acclaimed as a classic by statisticians and behavioral scientists alike. Written by William Sealy Gosset under the pseudonym "Student," its publication paved the way for the statistical era that continues today, one focused on how best to draw inferences about large populations from small samples of data.

## Gosset and "Student"

Schooled in mathematics and chemistry, Gosset was hired by Arthur Guinness, Son, & Co., Ltd. to apply recent innovations in the field of statistics to the business of brewing beer. As a brewer, Gosset analyzed how agricultural and brewing parameters (e.g., the type of barley used) affected crop yields and, in his words, the "behavior of beer." Because of the cost and time associated with growing crops and brewing beer, Gosset and his fellow "experimental" brewers could not afford to gather the large amounts of data typically gathered by statisticians of their era. Statisticians, however, had not yet developed accurate inferential methods for working with small samples of data, requiring Gosset to develop methods of his own. With the approval of his employer, Gosset spent a year (1906–1907) in Karl Pearson's biometric laboratory, developing "The Probable Error of a Mean" as well as "Probable Error of a Correlation Coefficient."

The most immediately striking aspect of "The Probable Error of a Mean" is its pseudonymous author: "Student." Why would a statistician require anonymity? The answer to this question came publicly in 1930, when fellow statistician Harold Hotelling revealed that "Student" was Gosset, and that his anonymity came at the request of his employer, a "large Dublin Brewery." At the time, Guinness considered its use of statistics a trade secret and forbade its employees from publishing their work. Only after negotiations with his supervisors was Gosset able to publish his work, agreeing to neither use his real name nor publish proprietary data.

## The Problem: Estimating Sampling Error

As its title implies, "The Probable Error of a Mean" focuses primarily on determining the likelihood that a sample mean approximates the mean of the population from which it was drawn. The "probable error" of a mean, like its standard error, is a specific estimate of the dispersion of its sampling

distribution and was used commonly at the start of the 20th century. Estimating this dispersion was then, and remains today, a foundational step of statistical inference: To draw an inference about a population parameter from a sampled mean (or, in the case of null hypothesis significance testing, infer the probability that a certain population would yield a sampled mean as extreme as the obtained value), one must first specify the sampling distribution of the mean. The Central Limit Theorem provides the basis for parametrically specifying this sampling distribution, but does so in terms of population variance. In nearly all research, however, both population mean and variance are unknown. To specify the sampling distribution of the mean, therefore, researchers must use the sample variance.

Gosset confronted this problem with using sample variance to estimate the sampling distribution of the mean, namely, that there is error associated with sample variance. And because the sampling distribution of the variance is positively skewed, this error is more likely to result in the underestimation than the overestimation of population variance (even when using an unbiased estimator of population variance). Furthermore, this error, like the error associated with sampled means, increases as sample size decreases, presenting a particular (and arguably exclusive) problem for small sample researchers such as Gosset. To draw inferences about population means from sampled data, Gosset could not—as large-sample researchers did—simply calculate a standard $z$ statistic and rely on a unit normal table to find the corresponding $p$ values. The unit normal table does not account for either the estimation of population variance or the fact that the error in this estimate depends on sample size. This limitation inspired Gosset to write "The Probable Error of a Mean" in a self-described effort to (a) determine at what point sample sizes become so small that the above method of normal approximation becomes invalid and (b) develop a set of valid probability tables for small sample sizes.

## The Solution: $z$

To accomplish these twin goals, Gosset derived the sampling distribution of a new statistic he called $z$. He defined $z$ as the deviation of the mean of a sample ($\overline{X}$) from the mean of a population ($u$) divided by the standard deviation of the sample ($s$), or $(\overline{X} - u)/s$. In his original paper, Gosset calculated $s$ with the denominator $n$ (leading to a biased estimate of population variance, $s^2$) rather than the unbiased $n - 1$, likely in response to Pearson's famous attitude that "only naughty brewers take $n$ so small that the difference is not of the order of the probable error!" To determine the sampling distribution of $z$, Gosset first needed to determine the sampling distribution of $s$. To do so, he derived the first four moments of $s^2$, which allowed him to make an informed guess concerning its distribution (and the distribution of $s$). Next, he demonstrated that $\overline{X}$ and $s$ were uncorrelated, presumably in an effort to show their independence. This independence—in conjunction with equations to describe the distribution of $s$—allowed Gosset to derive the distribution of $z$.

This first portion of "The Probable Error of a Mean" is noteworthy for its speculative, incomplete, and yet ultimately correct conclusions. Gosset failed to offer a formal mathematical derivation for the sampling distribution of $s$, despite the fact that, unbeknownst to him, such a proof had been published 30 years earlier by the German statistician Friedrich Robert Helmert. Nor was Gosset able to prove that the sampling distributions of $s^2$ and $\overline{X}$ were completely independent of each other. Nevertheless, Gosset was correct on both counts, as well as his ensuing derivation of the sampling distribution of $z$, leading many to note that his statistical intuition more than compensated for his admitted mathematical shortcomings.

## Pioneering Use of Simulation

"The Probable Error of a Mean" documents more than Gosset's informed speculation, however; it presents one of the first examples of simulation in the field of statistics. Gosset used simulation to estimate the sampling distribution of $z$ nonparametrically, and then compared this result to his parametrically derived distribution. Concordance between the two sampling distributions, he argued, would confirm the validity of his parametric equations.

To conduct his simulation, he relied on a biometric database of height and finger measurements

collected by British police from 3,000 incarcerated criminals; this database served as his statistical population. Gosset randomly ordered the data—written individually on pieces of cardboard—then segregated them into 750 samples of 4 measurements each (i.e., $n = 4$). For every sample, he calculated $z$ for height and finger length, then compared these two $z$ distributions with the curves he expected from his parametric equations. In both cases, the empirical and theoretical distributions did not differ significantly, thus offering evidence that Gosset's preceding equations were correct.

## Tables and Examples

Gosset dedicated the final portion of "The Probable Error of a Mean" to tabled probability values for $z$ and illustrative examples of their implementation. To construct the tables, he integrated over the $z$ distributions (for sample sizes of 4–10) to calculate the probability of obtaining certain $z$ values or smaller. For purposes of comparison, he also provided the $p$ values obtained via the normal approximation to reveal the degree of error in such approximation. The cumbersome nature of these calculations deterred Gosset from providing a more extensive table.

In further testament to his applied perspective, Gosset concluded the main text of "The Probable Error of a Mean" by applying his statistical innovation to four sets of actual experimental data. In the first and most famous example, Gosset analyzed data from a 1904 experiment that examined the soporific effects of two different drugs. In this experiment, researchers had measured how long patients ($n = 15$) slept after treatment with each of two drugs and a drug-free baseline. To determine whether the drugs helped patients sleep, Gosset tested the mean change in sleep for each of the drug conditions (compared to the baseline) against a null (i.e., zero) population mean. To test whether one drug was more effective than the other drug, he tested the mean difference in their change values against a null population mean. All three of these tests—as well as the tests used in the three subsequent examples—correspond to modern-day one-sample $t$ tests (or equivalent paired $t$ tests).

## Postscript: From $z$ to $t$

With few exceptions over nearly 20 years following its publication, "The Probable Error of a Mean" was neither celebrated nor appreciated. In fact, when providing an expanded copy of the Student $z$ tables to then little-known statistician Ronald Fisher in 1922, Gosset remarked that Fisher was "the only man that's ever likely to use them!" Fisher ultimately disproved this gloomy prediction by championing Gosset's work and literally transforming it into a foundation of modern statistical practice.

Fisher's contribution to Gosset's statistics was threefold. First, in 1912 and at the young age of 22, he used complex $n$-dimensional geometry (that neither Gosset nor Pearson could understand) to prove Gosset's equations for the $z$ distribution. Second, he extended and embedded Gosset's work into a unified framework for testing the significance of means, mean differences, correlation coefficients, and regression coefficients. In the process of achieving this unified framework (based centrally on the concept of degrees of freedom), Fisher made his third contribution to Gosset's work: He multiplied $z$ by $\sqrt{n-1}$, transforming it into the famous $t$ statistic that now inhabits every introductory statistics textbook.

During Fisher's popularization, revision, and extension of the work featured in "The Probable Error of a Mean," he corresponded closely with Gosset. In fact, Gosset is responsible for naming the $t$ statistic, as well as calculating a set of probability tables for the new $t$ distributions. Despite Gosset's view of himself as a humble brewer, Fisher considered him a statistical pioneer whose work had not yet received the recognition it deserved.

## Historical Impact

The world of research has changed greatly in a century, from a time when only "naughty brewers" gathered data from sample sizes not measured in hundreds, to an era characterized by small sample research. "The Probable Error of a Mean" marked the beginning of serious statistical inquiry into small sample inference, and its contents today underlie behavioral science's most frequently used statistical tests. Gosset's efforts to derive an exact

test of statistical significance for such samples (as opposed to one based on a normal approximation) may have lacked mathematical completeness, but their relevance, correctness, and timeliness shaped scientific history.

*Samuel T. Moulton*

***See also*** Central Limit Theorem; Distribution; Nonparametric Statistics; Sampling Distributions; Sampling Error; Standard Error of the Mean; Student's *t* Test; *t* Test, Independent Samples; *t* Test, One Sample; *t* Test, Paired Samples

**Further Readings**

Boland, P. J. (1984). A biographical glimpse of William Sealy Gosset. *American Statistician, 38,* 179–183.
Box, J. F. (1981). Gosset, Fisher, and the *t*-distribution. *American Statistician, 35,* 61–66.
Eisenhart, C. (1979). Transition from Student's *z* to Student's *t*. *American Statistician, 33,* 6–10.
Fisher, R. A. (1925). Applications of 'Student's' distribution. *Metron, 5,* 90–104.
Hanley, J. A., Julien, M., & Moodie, E. E. M. (2008). Student's *z, t,* and *s:* What if Gosset had *R? American Statistician, 62,* 64–69.
Lehmann, E. L. (1999). "Student" and small-sample theory. *Statistical Science, 14,* 418–426.
Pearson, E. S. (1939). 'Student' as a statistician. *Biometrika, 30,* 210–250.
Pearson, E. S. (1968). Studies in history of probability and statistics XX: Some early correspondence between W. S. Gosset, R. A. Fisher, and K. Pearson with note and comments. *Biometrika, 55,* 445–457.
Zabell, S. L. (2008). On Student's 1908 article—"The Probable Error of a Mean." *Journal of the American Statistical Association, 103,* 1–7.

# Propensity Score Analysis

Propensity score analysis is a technique for estimating the causal effect of a treatment in an observational study. Although randomized experiments are the ideal method for estimating the causal effect of a treatment—because randomization ensures that, on average, the distribution of both observed and unobserved characteristics are the same for treated and untreated units—there are many cases where randomized experiments are unethical or impractical. In an observational study, unlike in a randomized experiment, the researcher has no control over treatment assignment. As a result, due to self-selection into treatment or other nonrandom aspects of the treatment assignment, there may be systematic differences between the treated and control groups that can bias the estimate of treatment effects. Using propensity scores to match treated units to similar control units is one way to adjust for observed differences between the two groups and thereby reduce this selection bias in the treatment effect estimate. Bias may not be completely eliminated because propensity score analysis will not adjust for unobserved differences between the two groups, except to the extent that observed variables are correlated with these unobserved variables; however, this is a limitation of any nonrandomized study.

This entry describes the theory underlying this analysis. Next, the entry presents the steps involved in implementing propensity score analysis. The entry concludes with a brief discussion of alternate uses for propensity scores.

## Underlying Theory

Formally, the propensity score is defined as the conditional probability of receiving treatment given a set of observed covariates. For simplicity, treatment will be assumed to be a binary variable (i.e., treatment or control). The majority of propensity score research addresses this binary treatment case; however, the propensity score method has been extended to treatments with multiple doses. The propensity score is a one-number summary of the multivariate information in the covariates that are related to treatment assignment. More specifically, the propensity score is a balancing score in the sense that matching on the propensity score creates pairs or subclasses within which the distribution of observed covariates is the same, on average, for both treated and control groups. In other words, the covariate distributions are "balanced" between the treated and control groups within these subclasses. Under two key assumptions—strongly ignorable treatment assignment and stable unit treatment value assumption (SUTVA)—an unbiased estimate of the average treatment effect at a specific propensity score value is given by the difference between the treated and

control means for all units with that value of the propensity score. Averaging these differences across all propensity score values in the population yields an average overall treatment effect. The strongly ignorable treatment assumption implies that each unit has a nonzero probability of receiving treatment and that all relevant covariates have been included in the propensity score model so that treatment assignment is unconfounded with the covariates (sometimes referred to as "unconfoundedness" or "no hidden bias" or "selection on observables"). The SUTVA assumption implies that there is only one version of each treatment for each unit and that the treatment assignment of one unit does not affect the outcome of any other units ("no interference between units").

Of course, true propensity scores are never known and must be estimated. Estimated propensity scores can be used to create matched pairs or groups of treated and control units with similar propensity score values. Estimates of treatment effects are obtained by comparing treated and control units within these matched pairs or groups.

## Implementation of a Propensity Score Analysis

### *Select the Covariates*

In order to adjust for all possible systematic differences between the treated and control groups, common advice is to include as many covariates as possible in the propensity score model. More specifically, any covariates that are thought to be related to treatment assignment and the outcomes of interest should be included. Care must be taken, however, to include only variables that are not affected by treatment assignment. Including variables that are affected by the treatment can bias the treatment effect estimates.

### *Measure the Initial Imbalance in the Covariates*

Calculating a baseline measure of covariate imbalance is useful as a benchmark to judge how well the propensity score matching method improved balance. Many researchers use significance tests to test for balance in the covariates. For example, a two-sample $t$ test can be used to test the difference in the mean of a continuous covariate between the treated and control groups, and a two-sample test of proportions can be used for a binary covariate. However, this has been recently criticized as being sensitive to changes not only in balance but also in sample size. For example, some observations may be dropped during the matching process if good matches cannot be found for them, thereby reducing the sample size and power of the test. Although the best way to measure balance is still being debated in the statistical literature, one of the currently suggested alternatives to significance tests is to measure the standardized difference between the treated and control groups with a difference of .1 (as a rule of thumb) indicating substantial imbalance. Different versions of the fomula for calculating the absolute standardized difference appear in the literature. Two versions of the fomula for continuous covariates are

$$d = \frac{\left(\overline{X}_{\text{treatment}} - \overline{X}_{\text{control}}\right)}{\sqrt{\frac{S^2_{\text{treatment}} + S^2_{\text{control}}}{2}}}$$

or

$$d = \frac{\left(\overline{X}_{\text{treatment}} - \overline{X}_{\text{control}}\right)}{S_{\text{treatment}}}.$$

Note that the key features of standardized difference (which is true for both fomulas) are that it does not depend on the unit of measurement and, unlike a $t$ statistic, is not influenced by sample size.

### *Estimate the Propensity Score*

The most common method to estimate the propensity score is via logistic regression, but it is possible to use other methods, such as probit regression or even classification trees. Assuming a binary treatment, the logistic regression propensity score model can be specified as

$$\log\left(\frac{P_i}{1 - P_i}\right) = \alpha + \beta \mathbf{X}_i,$$

where $P_i$ is the probability that unit $i$ receives treatment (as opposed to control) and $\mathbf{X}_i$ is a vector of covariates and any relevant interactions thereof. This model yields $\widehat{P}_i$, the estimated propensity score, for each unit. The quality of the estimated propensity score model should be judged in terms of the balance that is achieved after matching on

the propensity scores. The usual goodness-of-fit measures (such as the area under the Receiver Operating Curve) and the statistical significance of the coefficients are not useful here. Similarly, because the goal of the propensity score model is to produce balanced subclasses of treated and control units, it is not necessary to search for a parsimonious model (unless a small sample size forces the analyst to limit the number of covariates in the propensity score model).

### *Check for Overlap in the Propensity Score Distributions*

A first step in matching or subclassifying on the estimated propensity score is to confirm that there is an adequate overlap in the distribution of estimated propensity scores (or "common support") for treated and control units to make reliable treatment effect estimates. This can be done by graphing overlapping histograms or by tallying sample sizes separately by treatment group within quintiles (for example) of the estimated propensity scores. If there is no overlap or minimal overlap in the distribution of propensity scores across treatment groups, then any estimates of the treatment effect will rely on model extrapolation. For example, in the extreme (and simplistic) case that all treated units are female and all control units are male, no treatment effect estimates can be made without making some assumptions about the relationship between males and females and the outcome. In a less extreme case, one may find that treatment effects can be estimated only within a certain subset of the population. For example, if the estimated propensity scores for the control group range between .10 and .65, and the estimated propensity scores for the treatment group range between .45 and .95, then one can estimate only the treatment effects for the subset of the population with estimated propensity scores between .45 and .65. In the range between .10 and .45, there are no treated units to compare to the control units, and in the range between .65 and .95, there are no control units to compare to the observed treated units. This diagnostic property of propensity scores has been cited as one of the advantages of the propensity score method over traditional methods such as linear regression.

At this point, the analyst must decide what to do with the overlap information. In some cases, the analyst may decide to restrict all further analysis to the region of propensity score overlap. In other cases, the analyst may find that there are only a few cases in the nonoverlapping tail area, and keeping them in the analysis does not adversely affect the balance in the covariates or the quality of the matches obtained.

### *Match or Subclassify the Estimated Propensity Score*

There are many propensity score matching methods. Without going into the details of the specific implementation of any particular matching method, the main matching issues are discussed here. User-written matching programs are available for Stata, R, and SAS.

*Size of the Matched Groups.* Nearest neighbor matching selects $k$ matched controls for each treated unit. The simplest and most common variation of this method uses $k = 1$ for matched pairs of treated and control units. Values of $k$ larger than 1 can be used when there are large numbers of control units relative to treated units. Also, $k$ can vary across treated units. Matching each treated unit to multiple control units can reduce sampling variance in the treatment effect estimates because of the larger matched sample size. On the other hand, because these extra matches are further away from the treated unit than the first closest match (unless the controls have the exact same covariate values), using multiple controls for each treated unit will increase the bias in our treatment effect estimates. Using multiple controls can also be more costly if, for example, additional data need to be gathered from the selected matches.

*Matching Algorithm.* The simplest way to select matches is to sequentially select the closest control unit for each treated unit. This type of "greedy" algorithm finds different matches depending on how the treated units are sorted because matches are selected without regard as to whether the untreated unit would better serve as a match to a different treated unit yet to be encountered in the list. This may have a substantial effect on the quality of the matches when the overlap in the propensity score distributions of the treated and

control units is small, but may not have much of an effect on the matches when there is substantial overlap in the propensity score distributions. A more sophisticated "optimal" matching algorithm finds the set of matches that minimizes a global measure of distance across all matches.

*Matching With or Without Replacement.* In with-replacement matching, controls can be used as matches for more than one treated unit, as opposed to without-replacement matching, where each unit can be used in only one match. Matching with replacement can be useful when there are few control units that are comparable to treated units. Also, when matching with replacement, the order in which matches are found does not matter. However, when matching is done with replacement, it is possible that subsequent treatment effect estimates are based on just a small number of controls. The number of times each control is used as a match should be monitored and possibly controlled.

*Matching for Key Covariates.* When it is important to find matches that are particularly close on a few key covariates, propensity score matching techniques can be extended to Mahalanobis metric matching within propensity score calipers. In this method, the closest control unit, in terms of Mahalanobis distance on the key covariates, within a prespecified distance of the treated unit's propensity score (the propensity score "caliper") is selected as a match.

*Subclassification.* Subclassification forms groups of units with similar propensity scores by stratifying on values of the propensity score. For example, observations can be divided into subclasses (or strata) based on the quintiles of the estimated propensity score (this will create five groups of equal total size), or, if the effect on the treated group is of particular interest, based on the quintiles of the estimated propensity score in the treated group (this will create groups with equal numbers of treated units). Typically, five subclasses are used; however, if sample size permits, a larger number of subclasses can be used. With more subclasses, treated and control units will be more similar within each class, but using more subclasses can stretch the sample too thin, leaving some subclasses with few treated or control units, thereby decreasing the precision of the treatment effect estimates.

Subclassification has the advantage over matching of being very easy to implement (although matching software is getting easier to obtain and use) and of using all the data, whereas matching may throw away data from unmatched units. This may be a particularly important issue when generalizing to a larger population (e.g., via survey sampling weights).

Subclassification can be viewed as a very coarse version of matching. To obtain more of the benefits of matching, subclassification has been extended to "full matching," which creates matched sets containing either (a) one treated unit and one or more controls or (b) one control unit and one or more treated units.

### *Check for Balance Within the Propensity Score Groups*

To check whether matching or subclassifying on the propensity score did indeed remove the initial imbalance in the covariates, it is necessary to assess the balance within the matched groups. As stated before, experts disagree on the best method to assess this balance. Some experts recommend significance tests. For example, *t* tests of the difference in covariate means between the matched treated units and the matched control units, or, in the case of subclassification, a two-way analysis of variance with the covariate as the dependent variable and treatment indicator and propensity score subclass and their interaction as predictors, can be used as a guide to indicate remaining covariate imbalance. Another recommendation is to calculate standardized differences, keeping the denominator the same as in the prematch balance checks. More specifically, the measure of postmatch standardized difference uses the standard deviation of the entire treatment group (and the entire control group if using the first version of the fomula) in the denominator, even if some units have been dropped in the matching process because of the unavailability of a good match in the opposite treatment group.

If covariate imbalance remains at this step, the propensity score model can be reestimated, adding interaction terms or nonlinear functions (e.g.,

quadratic) of imbalanced covariates to try to improve the resulting balance.

### *Estimate the Treatment Effect*

When propensity score matches balance all the covariates to the analyst's satisfaction, the simplest way to estimate the treatment effect is to calculate the difference in average outcome between the treated and control units within a matched group and then average these differences across all the groups. When subclassification is used, using a weighted average based on the number of treated units in each group will yield an average treatment effect on the treated units, and using a weighted average based on the total number of units in each subclass will yield an average treatment effect for the overall population.

When covariate imbalance remains after propensity score matching, despite repeated modeling attempts to improve the propensity score estimates, a regression-type adjustment can be used to adjust for the remaining differences between groups. For example, the matched treated and control units can be used in a multiple regression of the outcome on the imbalanced covariates and an indicator for treatment. In the case of subclassification, this regression can be carried out within strata and the estimated treatment effects (the coefficient of the treatment indicator) can be averaged across strata.

When with-replacement matching has been used, or when treated units have been matched to different numbers of controls, this must be taken into account when estimating the treatment effect. Usually, the analysis uses the combined sample of matched pairs, with units that were selected as a match more than once receiving a larger weight (e.g., if a control was used twice as a match, then it would receive a weight of two).

Variance estimation for propensity score estimates of treatment effects is still a topic of open research. Experts are debating whether uncertainty in the propensity score estimation or the effect of sampling variability in the matching process needs to be taken into account. Therefore, any variance estimates obtained from propensity-matched data should be viewed as only approximate.

### Alternative Uses of Propensity Scores

Propensity scores have also been used as covariates in models such as linear regression, and as weights.

*Elaine Zanutto*

***See also*** Logistic Regression; Matching; Multiple Regression; Randomization Tests

#### Further Readings

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys, 22,* 31–72.

D'Agostino, R. B. (1998). Tutorial in biostatistics: Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine, 17,* 2265–2281.

Ho, D., Imai, K., King, G., & Stuart, E. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis, 15*(3), 199–236.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41–55.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79,* 516–524.

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine, 127*(Part 2), 757–763.

Rubin, D. B. (2006). *Matched sampling for causal effects.* Cambridge, UK: Cambridge University Press.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin.

Stuart, E. A., & Rubin, D. B. (2007). Best practices in quasi-experimental designs: Matching methods for causal inference. In J. Osborne (Ed.), *Best practices in quantitative social science* (pp. 155–176). Thousand Oaks, CA: Sage.

## Proportional Sampling

Proportional sampling is a method of sampling in which the investigator divides a finite population

into subpopulations and then applies random sampling techniques to each subpopulation. Proportional sampling is similar to proportional allocation in finite population sampling, but in a different context, it also refers to other survey sampling situations. For a finite population with population size $N$, the population is divided into $H$ strata (subpopulations) according to certain attributes. The size of the $h$th stratum is denoted as $N_h$ and $\sum_{h=1}^{H} N_h = N$. Proportional sampling refers to a design with total sample size $n$ such that

$$n_h = n\frac{N_h}{N}$$

and

$$\sum_{h=1}^{H} n_h = n.$$

Following that, a simple random sample with sample size $n_h$ would be selected within each stratum. This is essentially the same as the stratified random sampling design with proportional allocation, and the inference of the population quantity of interest can be established accordingly. $n_h$ usually would not be an integer, so the closest integer can be used. For example, in an opinion survey, an investigator would usually prefer a sample with the sample structure as close as possible to the population structure with respect to different population characteristics, such as age, gender, race, and so forth. Suppose the population structure of gender by age for people more than 20 years of age is as summarized in Table 1, and the total sample size $n$ is determined to be 1,068. The sample sizes of each stratum can be determined as shown in Table 2.

**Table 1** Population Structure

| | | *Age* | | | | |
|---|---|---|---|---|---|---|
| | | *20–29* | *30–39* | *40–49* | *50–59* | *≥ 60* |
| *Gender* | *Male* | 22,306 | 25,264 | 28,231 | 22,306 | 17,350 |
| | *Female* | 24,785 | 27,646 | 29,742 | 22,906 | 24,785 |

**Table 2** Sample Sizes

| | | *Age* | | | | |
|---|---|---|---|---|---|---|
| | | *20–29* | *30–39* | *40–49* | *50–59* | *≥ 60* |
| *Gender* | *Male* | 97 | 110 | 123 | 97 | 76 |
| | *Female* | 108 | 120 | 129 | 100 | 108 |

For example, the sample size of females between the ages of 30 and 39 can be determined by

$$1068 \times \frac{27646}{245321} = 120.3563 \cong 120,$$

where 245,321 is the total population size.

For other sampling survey situations—for instance, the selection of sites to monitor the air population level in a region or time points to observe a production process for quality control purposes during a period of operation—the study region can be viewed as a compact region and the number of possible sampling units is infinite. Suppose that the study region $D$ is divided into $H$ disjoint domains $D_h$, $h = 1, \ldots, H$, according to the practical or natural demands; that is,

$$D = \bigcup_{h=1}^{H} D_h, D_h \bigcup D_{h'} = \varphi, \forall h \neq h'.$$

For example, for ecological research on Florida alligators, the study region can be a collection of several lakes, and the domains are the lakes. To study the ecosystem in a certain lake, the investigator might divide the study region, a three-dimensional compact space of the lake, by water depth. The investigator would like to allocate the sample sizes proportional to the area $D_h$ to ensure a fair devotion of sampling effort to each domain. Proportional sampling can also be used such that

$$n_h = n \cdot \frac{\text{area of } D_h}{\text{area of } D}$$

and

$$\sum_{h=1}^{H} n_h = n.$$

For example, an investigator would like to take samples of water from several lakes for a water pollution study. Suppose that there are five lakes,

**Table 3** Areas of Lakes in Acres

| | *Lakes* | | | | |
|---|---|---|---|---|---|
| | *A* | *B* | *C* | *D* | *E* |
| *Area* | 36.7 | 48.2 | 15.3 | 26.1 | 79.6 |

**Table 4** Sample Sizes in Each Lake

| | *Lakes* | | | | |
|---|---|---|---|---|---|
| | *A* | *B* | *C* | *D* | *E* |
| *Area* | 27 | 35 | 11 | 19 | 58 |

denoted from A to E, and the areas of these lakes are summarized in Table 3. If the sample size is determined to be 150, then the sample sizes in each lake are as summarized in Table 4.

For example, the sample size in Lake C can be determined as

$$150 \times \frac{15.3}{250.9} = 11.14619 \cong 11,$$

where

$$250.9 = 36.7 + 48.2 + 15.3 + 26.1 + 79.6$$

is the total area of the study region.

The principle of proportional sampling is widely used in practical sampling survey cases, either to determine the sample sizes of each domain or to evenly spread sampling units over the study region. Numerous researchers recommend proportional sampling because it is relatively more efficient than the uniform sampling or simple random sampling on the whole study region. To name a few, José Schoereder and colleagues described the application of proportional sampling in an ecological survey. Saul Blumenthal discussed proportional sampling on a time interval, in which the study region is a one-dimensional compact space. T. Y. Chen and colleagues studied the performance of proportional sampling on software testing and found it is a better choice than the random testing.

*Chang-Tai Chao*

*See also* Random Sampling; Sample Size; Stratified Sampling

**Further Readings**

Blumenthal, S. (1967). Proportional sampling in life length studies. *Technometrics, 9*(2), 205–218.

Chen, T. Y., Tse, T. H., & Yu, Y. T. (2001). Proportional sampling strategy: A compendium and some insights. *Journal of System and Software, 58*(1), 65–81.

Schoereder, J. H., Galbiati, C., Ribas, C. R., Sobrinho, T. G., Sperber, C. F., DeSouza, O., & Lopes-Andrade, C. (2004). Should we use proportional sampling for species-area studies? *Journal of Biogeography, 31,* 1219–1226.

Thompson, S. K. (2002). *Sampling.* New York: Wiley.

# PROPOSAL

A program research plan, or proposal, is a comprehensive description of a potential study that is intended to examine a phenomenon in question. The main purpose of research is the increase of knowledge, either in a specific disciplinary field or in the practice of a professional field. Research is usually associated with a planned, systematic investigation that produces facts and ideas that contribute to human thought, reflection, and action. Because there is no single way in which to approach research, an inquirer must weigh many factors in designing a program research plan.

Throughout the centuries, basic research structures have developed as models for each subsequent generation of knowledge-seekers. One who chooses to research a particular concept or phenomenon within one's field may look to the historic development of research methodology in that particular field to guide him or her in designing an appropriate research protocol. Regardless of the field of study, the potential researcher will find that scholarly investigation is guided by two overarching systems of inquiry—*quantitative* and *qualitative* research methodologies. Both are viable avenues of study that address inquiry in multiple disciplines. However, they are guided by different principles and are, therefore, best applied when matched to the characteristics of the problem that is to be studied.

Generally, postsecondary institutions of education require that a proposal or prospectus be submitted, reviewed, and approved prior to emoverlineking upon a thesis or dissertation study. This requires that the researcher complete a preliminary review of related literature and fomulate a hypothesis or research question that will provide focus for the study. Following these initial steps, the remainder of the program research study is designed and defended as part of the approval process. This plan must also be submitted to the college or university's research department for approval through the institutional review board. The approvals process has many steps, but a fully developed, well-designed plan saves time and effort in the long run; provides structure for the study; and normally produces a higher quality of research.

Although the exact titles, formatting, and subsections may vary based upon the research design and researcher preferences, research plans, whether quantitative or qualitative in nature, have several elements that are commonly included in their designs. Both usually include introductory components that present the phenomenon to be studied, a review of relevant literature, descriptions of the research design and procedures, and discussion of the data analysis methods to be used within the study.

## Quantitative Research Versus Qualitative Research

Quantitative research methodologies are most often associated with scientific investigation of quantifiable properties and their relationships. It uses mathematical models, theories, and hypotheses to measure and portray the empirical associations found in natural phenomena. It is frequently the research methodology employed for inquiry in natural science and social science fields such as physics, biology, psychology, and sociology. Characteristically, quantitative methods use large, random samples from which to gather data, and they require that the researcher remain in a neutral, noninteractive role so as to remove any bias that could affect the outcome of the study. Data are often numerical in nature, collected using reliable and valid tools and methods, and analyzed using statistical techniques. Results and findings are methodically presented following statistical data analyses, focusing on deductive, logical reasoning.

Qualitative research methodologies provide a means of gaining deeper understandings of human behaviors and the factors that influence those behaviors. The nature of human interactions is dynamic and may be reshaped by myriad factors. Therefore, qualitative methodology offers multiple means of inquiry that are flexible and adaptable within a more fluid research protocol. In essence, the researcher does not present a typical hypothesis at the onset of the study. Instead, he or she develops the research questions and focus during the study. The questions and focus of the study may change as the researcher investigates and collects data for the project. The broad range of human behaviors dictates that qualitative research use smaller samples in order to limit the scope of a study and allow the inquiry to focus more deeply on the phenomena to be examined. Unlike quantitative research, qualitative research is often subjective in nature, relying on interaction between the researcher and study participants. Qualitative data are characterized by their descriptive qualities, collected through personal accounts and interviews. Instead of using statistical analysis techniques, the qualitative researcher investigates patterns and common themes within the data. Study results are written in narrative form, using inductive reasoning to present findings.

Both research methodologies offer advantages and disadvantages; therefore, the researcher must look to the focus and purpose of his or her inquiry to determine which methodology is the better match for a proposed study. Some of the basic features of qualitative and quantitative research are highlighted in Table 1.

## Quantitative Proposed Program

The choice of a quantitative research design requires the researcher to fomulate an initial, specific hypothesis that will serve to guide the study. Thought and planning must be incorporated into the development of instruments and methods for measurement. Variables must be identified and their control and manipulation must be factored into the research design.

Selection of a quantitative research methodology brings with it structure and format that are commonly accepted as the scholarly progression to conduct such a study. Table 2 provides an outline

**Table 1** Basic Features of Quantitative and Qualitative Research Methodologies

| *Qualitative* | *Quantitative* |
|---|---|
| Generates detailed descriptions portrayed in creative, narrative forms | Generates models, theories, and hypotheses portrayed in mathematical expressions |
| Role of the researcher is interactive and may include participation in the study | Role of the researcher is a neutral stance in order to eliminate any bias |
| No hypothesis—the researcher may begin a study with a concept, but without knowing in advance what he or she is specifically looking for | Hypothesis is stated—the researcher clearly knows in advance what he or she is looking for and seeks to prove or disprove the problem |
| The design may be flexible and aspects may be adapted and changed as the study progresses | All aspects of the design are clearly structured prior to the start of the study and remain intact |
| Researcher is the main data collection tool—use of interviews, observations, and so on rest with the researcher and require a departure from value neutrality | Researcher uses tools or instruments to collect numerical data—questionnaires, assessments, scales, and so on provide data through a neutral stance |
| Smaller samples of data are used because of the greater bulk of data generated | Large samples are used to provide maximum insight into numerical data |
| Data collection and analysis require extensive expenditure of time | Data collection and analysis are more efficient, requiring less time |
| Data analysis uses both subjective and objective means to determine findings | Data analysis is determined objectively through the use of mathematical calculations |
| Findings are presented in the form of words, pictures, or other objects | Findings are presented in comparative mathematical expressions |
| May be more specific rather than generalizable | Generalizable, but has less contextual detail |

of the elements that are contained within that academically sanctioned format.

### *Approaches to a Quantitative Research Plan*

Quantitative research is the method of choice for a researcher who seeks to clarify phenomena through specifically designed and controlled data collection and analysis. Within that methodology lies a broad spectrum of approaches.

*Experimental Research*. Identifies at least one independent variable for interventions and is manipulated, whereas other, related variables are controlled; effects on dependent variables are observed (e.g., measuring student achievement levels by examining and controlling selected variables such as age, grade level, reading level, time, teacher characteristics, etc.); two subgroups fall within this category: true experimental and quasi-experimental research.

- *True Experimental Research*. Participants are randomly selected and placed into control and experimental groups prior to the intervention; the experimental group receives the intervention and the control group does not; random selection and placement of the participants allows the researcher to control for extraneous variables; this method provides for true causal relationships between independent and dependent variables.
- *Quasi-Experimental Research*. This type of research is used when the researcher cannot or should not randomly assign participants to control and experimental groups; lack of random

**Table 2** Organizational Format for a Quantitative Research Study

| Preliminary Introductory Pages |
|---|
| Title Page |
| Acknowledgments (optional) |
| Abstract—brief, single-page overview of the study |
| Table of Contents |
| List of Tables |
| List of Figures |
| **Chapter 1—Introduction** |
| Introductory Paragraph |
| Statement of the Problem |
| Purpose of the Study |
| Significance or Rationale for the Study |
| Limitations of the Study |
| Delimitations of the Study |
| Operational Definitions |
| Statement of the Hypotheses or Research Questions |
| Summary of Chapter (and overview of the study) |
| **Chapter 2—Review of Related Literature** |
| Introductory Paragraph (may include organization of the review) |
| Review of the Literature |
| Theoretical Framework |
| Summary of Chapter (one paragraph) |
| **Chapter 3—Methodology** |
| Introductory Paragraph |
| Review of Hypotheses and/or Research Questions |
| Design of the Study |
| Population and Sampling (also include sampling frame) |
| Instrumentation (validity and reliability) |
| Procedures of the Study |
| Pilot Study (if needed) |
| Data Analysis |
| Summary of Chapter (one paragraph) |
| **Chapter 4—Findings** |
| Introductory Paragraph |
| Results of Data Analysis |
| Findings by Research Questions or Hypotheses |
| Summary of Chapter (one paragraph) |
| **Chapter 5—Conclusions, Implementations, and Recommendations** |
| Introductory Paragraph |
| Conclusions of the Study |
| Generalizability of the Findings |
| Logistical Aspects of the Study and Some Lessons Learned |
| Implications for Policy and/or Practice |
| Recommendations for Future Research |
| Summary of Chapter (one paragraph) |
| **Appendixes** |
| Time Line for the Study |
| Consent Form, IRB Approval |
| Samples of Data Collection Instruments—Surveys, Questionnaires, etc. |

assignment limits the researcher's ability to control for extraneous variables; results should be interpreted cautiously; these types of designs cannot prove cause and effect, although they can show relationships between variables.

*Correlational Research*. Seeks to examine the degree of relationship that may exist between two or more variables in a single group of participants through collection of relevant data; this type of research cannot definitely identify cause-and-effect relationships because of a lack of random selection and assignment of participants, and because of a lack of variable manipulation; lack of randomization also limits the generalizability of results.

*Causal-Comparative Research*. Investigates the causes or consequences of existing differences in groups of individuals; may also be referred to as ex post facto research due to the nature of dealing with groups that have already been established (e.g., whether students whose parents have attended college enter postsecondary education at a higher rate than students whose parents who have not attended college); because the events have already occurred, variable manipulation is impossible; results must be interpreted with caution and generalizability is limited.

*Survey Research*. Collects descriptive data from a targeted population in order to determine the current status of one or more specific variables (e.g., the status of health care in a given community); questions are not standardized, which may lead to issues with instrument validity and reliability.

### *Components of a Quantitative Research Plan*

For a solid proposal, the researcher needs to write in such a way that the study can be envisioned by the readers. The proposal should be clear and concise, with attention to detail. The following subgroups represent the sections of a written plan for conducting quantitative research. Although the subtitles may not be the same for every proposal, this information serves as an outline for developing the plan of research. The researcher will need to determine the exact subgroups and titles based upon research methods and type of study being conducted.

#### Introduction and Statement of the Problem

When writing a quantitative research plan, the researcher must first set the scene for the project. The introductory section of the plan allows the researcher to discuss the framework for the project, as well as provide vital background information. The researcher wants to draw in the audience, creating interest and intrigue. In this section, the researcher discusses the topic that is the focus of the research, including its importance to the advancement of knowledge in the particular field of study.

#### Review of Related Literature

The goal of the literature review section is to link current research and information to the researcher's focus of study, providing a foundation for the project. The literature review should provide adequate background information, current trends in the field of study, and prior research that relates to the project, as well as any gaps in the literature. The literature review will assist the researcher in developing research questions and hypotheses by relating the review findings to the goals of the project. The researcher should aim at including the most current literature and research available, relating it to the study at hand. Ideally, the researcher should demonstrate that his or her research is theoretically based, yet unique in its approach and what it intends to prove.

#### Statement of the Hypothesis

As mentioned previously, current literature and research trends should help the researcher determine the basic research questions and goals of the study. The literature review should guide the researcher in fomulating the hypothesis or hypotheses for the project. The statement of the hypothesis in a quantitative research proposal should be succinct and include the variables upon which the study will focus. In the proposal stage of the study, it is not always necessary for the researcher to know what type of effect the variables will have on one another. However, it is important that the researcher define clearly and operationally any variables that will be manipulated in the proposed program of study. In addition, if several research questions and/or hypotheses will be examined, all of these should be listed as part of this section. The final statement of the hypothesis or hypotheses must be written in such a way that anyone reviewing the proposal knows exactly what will be studied.

#### Methodology

The methodology section of the proposed program of a quantitative study describes exactly how the research will be conducted. The researcher needs to include methods involved in selecting participants, any instruments that will be used in the project, and the general research design.

*Selecting Participants.* When fomulating a quantitative research proposal, the researcher needs to include a detailed section on how participants will be selected. The researcher must also describe pertinent demographic information regarding the selected participants and the overall pool from which the participants are being chosen. The researcher must include the type of sampling procedures that will be used to select the participants, as well as how many participants will be included in the project.

*Instruments.* The researcher should discuss any data collection tools that may be used in the study. Instruments must be described clearly, and the researcher should specify how each will be used. In this section, it is important to indicate which are existing instruments, as well as which instruments will be developed by the researcher. It is also appropriate to discuss validity and reliability of instruments, details of how the instruments will be administered and scored, and why

the specific instruments were selected or developed for the study. Because this is the proposal phase of the study, the researcher may not have all of the data collection tools selected or developed at the time of submission. In such cases, the researcher should, at a minimum, discuss the type of instruments that will be used, such as interview questions, pretests, and so on.

*Research Design.* The research design outlines the framework for how the study will be conducted, including the research goals and questions that will be answered through the study. The goal of quantitative research is to examine relationships among or between variables. Several designs accomplish this objective as mentioned previously, including experimental and quasi-experimental, correlational, and causal-comparative research. The intention is for the researcher to select the design that best suits the current study's objectives and goals.

#### Data Collection

Some researchers include the data collection section within the methods section, but it may also be discussed separately. Regardless of how the proposal is organized, it is important for the researcher to effectively share the steps of collecting data for the project. This section should include how the participants will be selected, as well as the details regarding the data collection instruments and any other pertinent information relating to the data. The researcher must specify who will administer each tool, as well as how and to whom the tool will be given for data collection purposes. The researcher may also give an estimated time line for when the data will be collected. If the study claims any assumptions, these should be detailed, as well as any limitations that may influence the data and study outcomes.

#### Data Analysis

Although the researcher may not yet know the results of data analysis, he or she should be prepared to discuss the methods of how the data will be analyzed. This includes any inferential and descriptive statistical testing and techniques. The types of statistical methods chosen should flow from both the research goals and objectives and the design of the study. A variety of techniques can be employed in a study, but they depend upon the type of participant sampling methods, research design, independent and dependent variables, and data that are collected. The researcher can strengthen the proposal if he or she determines the specific techniques to analyze the data.

#### Report of Findings

Once the data have been analyzed, the final step is to report the findings of the study. The researcher conveys whether or not the study provided enough evidence to accept the hypothesis as true, the extent of how the results can be generalized and to what populations, and any other findings that were discovered during the study. This section should also point out any limitations of the study and discuss ways that further research could strengthen the study's hypothesis. Finally, the researcher should note the significance that the study has with regard to furthering knowledge in the field.

## Qualitative Proposed Program

Qualitative research explores the *why* and *how* of a phenomenon and provides opportunities for a researcher to go beyond the quantitative questions of *how many* and *to what extent* an experience may occur. According to Norman Denzin and Yvonna Lincoln, a qualitative research project permits the researcher to depart from the value neutrality usually associated with scientific research and delve into the variant factors surrounding a phenomenon—natural settings, characteristics of participants, interactions, thoughts, and discoveries—that are brought to light through interaction with the research participants themselves.

Characteristically, qualitative methodology allows the researcher to collect data through participation, observation, interviews, and analysis of artifacts and documents. With this information, the researcher may then attempt to interpret or make sense of the phenomenon through the meanings that the participants bring to them. Reporting of findings is most frequently done in a more literary form rather than being graphed, charted, or put forth in mathematical terms.

Qualitative methods have their early roots in sociology and anthropology as researchers sought to investigate lived human experiences. Usage expanded to include other social and behavioral sciences. Currently, researchers in multiple fields such as education, health services, clinical research, women's studies, and societal reform movements use qualitative methods to gain greater insight into specific experiences or phenomena. As an example, a researcher studying the importance of mental attitude in the success of chemotherapy for specific cancer patients may, through qualitative methods, explore the various attitudes of the patient, his or her family and friends, and any other factors that produce positive and/or negative feelings regarding the cancer treatments. In educational settings, qualitative methods allow the researcher to take on the role of the researcher/practitioner, thus enabling a teacher to engage in deeper evaluation of instructional techniques, resulting in improvements in both teaching and learning within the classroom.

Qualitative research methodology also lends itself to a *mixed methods* approach that employs both qualitative and quantitative procedures to expand the depth and scope of a research study. This inherent flexibility allows room for human variance and, as such, is a characteristic that has made qualitative research the methodology of choice for a broad range of social, behavioral, and biomedical sciences.

### *Approaches to a Qualitative Research Plan*

Qualitative researchers have an array of approaches that may be used to design a research study. Selection of a specific approach provides a framework that serves to guide the structure and sequence of a qualitative research study. Specific methods of collecting and analyzing data within each approach are determined by the nature of the phenomenon and should be selected to provide consistency between the occurrence and the data derived through that occurrence.

A researcher who elects to pursue a qualitative study faces the task of determining the approach that will fulfill the research goals. Consideration must be given to the research question and the audience for whom the research is targeted. John Creswell identifies five general models that are usually associated with qualitative research approaches.

1. *Ethnography.* Studies an intact culture in its own setting over a prolonged time period through direct observation.
2. *Grounded theory.* Seeks to establish a generalized theory regarding an experience or process that is grounded in the perspectives of the research participants. Constant comparison of data collected in multiple stages is a primary characteristic of this approach.
3. *Case study.* Explores events, processes, or activities in depth through the perspective of one or more participants over a specified time period. Detailed data collection using a wide range of measures is a hallmark of this approach.
4. *Phenomenological study.* Seeks to gain understanding of the lived experiences of a small number of participants in order to shed light on the phenomenon of such experiences.
5. *Narrative research*. Relies on the stories of individual participants to form the basis of a chronological account that merges the views of the participants with the views of the researcher into a collaborative narrative.

Each of these qualitative approaches carries with it characteristics and procedures that can dictate whether or not that approach will be consistent with the problem or phenomenon to be studied. Matching the approach with the problem will produce the most accurate and informational results for the researcher.

### *Components of a Qualitative Proposed Program*

The researcher's choice of qualitative approach directs the design and development of the research study. Research questions, review of relevant literature, and all aspects of data collection and analysis are fomulated relative to the qualitative approach selected for the study. The following information provides the basic structure for writing the qualitative research proposal. However, it is important to keep in mind that the subgroups

may be slightly different from those presented here depending on the researcher, methods used, and type of study.

### Statement of the Problem

Qualitative research is often the method of choice for those who seek to examine societal issues, especially those that confront marginalized people, political issues, and issues that affect specific groups of people. An effective statement of the problem sets the stage for the research and provides continual focus as the study progresses.

**Table 3** Organizational Format for a Qualitative Research Study

| Preliminary Introductory Pages |
|---|
| Title Page |
| Acknowledgments (optional) |
| Abstract—brief, single-page overview of the study |
| Table of Contents |
| List of Tables |
| List of Figures |
| **Chapter 1—Introduction of the Study** |
| Introductory Paragraph |
| Describe the Purpose of the Research Study |
| Context of the Study—Frame the Study Within the Larger Theoretical, Policy, or Practical Problem |
| Determine the Purpose of the Study |
| Initial Research Questions Are Posed |
| Related Literature Is Presented to Help Frame the Research Questions |
| Delimitations and Limitations of the Study |
| Significance of the Study |
| Summary of Chapter (and overview of the study) |
| **Chapter 2—Research Procedures** |
| Introductory Paragraph |
| Overall Approach and Rationale of the Study |
| Site and Sample Selection |
| Define the Researcher's Role |
| Determine Data Collection Methods and Data Management Strategies |
| Design Data Analysis Techniques to Be Used |
| Consider Trustworthiness Features Such as Triangulation of Data |
| Observe Ethical Considerations as Related to Human Subjects and Confidentiality of Participants |
| Summary of Chapter (usually one paragraph) |
| **Chapter 3—Reporting Findings** |
| Introductory Paragraph |
| Narrative Format Rather Than Mathematical Representations |
| Contributions of the Research |
| Limitations of the Study |
| Provided to All Stakeholders |
| Summary of Chapter (usually one paragraph) |
| **Appendixes** |
| Time Line for the Study |
| Consent Form, IRB Approval |
| Samples of Data Collection Instruments—Surveys, Questionnaires, etc. |

### Review of Relevant Literature

The literature review defines the importance of a qualitative study. Through this review, a researcher can determine whether there is a gap in knowledge or if there are other relevant studies that could serve as the basis for further examination of a specific problem. A thorough review of relevant literature suggests direction for a potential study and helps the researcher to determine the research questions. Furthermore, it contributes to the fomulation of ideas throughout the study and serves to confirm the need for continued exploration of a phenomenon.

There is, however, another option within qualitative research methodology for examining relevant literature. According to Michael Quinn Patton, reviewing the literature prior to the study has the potential to predispose the researcher's thinking, which could diminish openness to developments that occur during the study. Therefore, an alternative approach is to conduct the literature review simultaneously with the study so that emerging data and relevant literature create a collaborative source that informs the researcher's perspective as it continually develops. Regardless of whether the researcher decides to examine the literature in the initial stages of a study or to review it throughout a study, the

literature review is an essential factor to position a qualitative study and provide focus for its merit.

### Organizational Format

Academic expectations and standards dictate that any research study be organized in a logical and scholarly fashion. A commonly accepted format for the organizational progression of the "pieces and parts" of a qualitative study includes classic research elements, but also allows expanded creativity because of the flexible nature of qualitative research. For example, emerging data may suggest that an additional data source be added or expanded. In like manner, data analysis may point the researcher to a different body of relevant literature that supports or disputes those themes and categories that emerge throughout the analysis process. However, an overall consistency is provided through use of those elements that are widely accepted as necessary parts of a qualitative study. A summary of the common elements found within the organizational format for a qualitative research study is delineated in Table 3.

## *Organizational Format for a Qualitative Research Study*

### Data Collection

Qualitative research studies use small, in-depth samples from which to derive data. The most frequently used data collection strategies include the following:

- *Interviews*. These may be done one-on-one or in a group setting either in person or via telephone; information may be audio- or videotaped for transcription analysis.
- *Observations*. These are direct, firsthand accounts as derived by the researcher in either a concealed or known role; they may also include the participant as an observer.
- *Documents*. These include the researcher's field notes, journals, or other printed materials such as newspapers, minutes of meetings, or other public documents.
- *Audiovisual materials*. Photographs, videotapes, or artwork are among materials used in data collection procedures.

Of these methods, interviews are the primary source of for production of qualitative data. According to Donald Polkinghorne, interviews allow the qualitative researcher to obtain first-person accounts of participants' experiences. John Creswell noted that interviews with participants provide historical information that enhances understanding. The researcher designs the interview protocol so as to elicit the information needed to enlighten the research questions that focus the study. The use of open-ended questions allows participants to respond at greater depth and opens the opportunities for the researcher to ask questions that clarify any comments revealed within the context of the interview.

### Data Analysis

Interpretive qualitative research designs have several key characteristics in common. The first is that the researcher strives to understand the meaning that people have constructed about an experience; he or she looks for a depth of understanding not for the future, but for the present situation, the "here and now" of a setting. The second key characteristic is that the researcher serves as the primary instrument to collect and analyze data. Because understanding the human experience is the goal of the research, this human mechanism is the most ideal means of collecting and analyzing data due to the flexibility, adaptiveness, and immediacy brought to the task by the researcher. This brings inherent biases, but another characteristic of such research is to identify and monitor these biases, thus including their influence on data collection and analysis rather than trying to eliminate them. Finally, data analysis in an interpretive qualitative research design is an inductive process. Data are richly descriptive and contribute significantly as the text is used to build concepts and theories rather than to deductively test hypotheses.

It is the task of the researcher to include in the research design the method of data analysis that is to be used. Commonly accepted procedures include phenomenological analysis, hermeneutic analysis, discourse analysis, and a variety of coding procedures. Validity and trustworthiness of findings may be achieved by using multiple data collection methods and triangulating the data gleaned from those processes.

### Reporting Findings

Findings derived from a qualitative research study are commonly presented in a narrative style. The flexibility that is the hallmark of qualitative research is also present when the researcher designs the manner for presentation of findings. Sharon Merriam noted that no standard format is required for reporting qualitative research, but that a diversity of styles is allowable with room for creativity. Narratives may be accompanied by commonly accepted methods such charts and graphs, or may be illustrated by photographs or drawings.

*Ernest W. Brewer and Nancy Headlee*

***See also*** Experimental Design; Ex Post Facto Study; Nonexperimental Design; Planning Research; Prospective Study; Qualitative Research; Quantitative Research

### Further Readings

Creswell, J. W. (2008). *Educational planning* (3rd ed.). Upper Saddle River, NJ: Pearson.
Denzin, N. K., & Lincoln, Y. S. (Eds.). (2005). *Handbook of qualitative research* (3rd ed.). Thousand Oaks, CA: Sage.
Fraenkel, J. R., & Wallen, N. E. (2009). *How to design and evaluate research in education* (7th ed.). New York: McGraw-Hill.
Gay, L. R., Mills, G. E., & Airasian, P. (2009). *Educational research: Competencies for analysis and applications*. Upper Saddle River, NJ: Pearson.
Merriam, S. B. (2002). *Qualitative research in practice: Examples for discussion and analysis*. San Francisco: Jossey-Bass.
Patton, M. Q. (2002). *Qualitative research and evaluation methods* (3rd ed.). Thousand Oaks, CA: Sage.

## Prospective Study

The term *prospective study* refers to a study design in which the documentation of the presence or absence of an exposure of interest is documented at a time period preceding the onset of the condition being studied. In epidemiology, such designs are often called *cohort studies*. Characteristic features of these designs include initial selection of study subjects at risk for a condition of interest, but free of disease at the outset. The occurrence/nonoccurrence of the condition is then assessed over a time period following recruitment. The time sequencing of the documentation of the exposure and outcome is an important feature and is related to a commonly cited form of evidence toward drawing causal inference about the relationship between the exposure and the outcome. The terms *prospective study* and *cohort study* also are usually used for observational or nonexperimental designs in which the investigator has not assigned study participants to the levels of exposure or intervention, distinguishing this design from prospectively conducted experiments or trials. In this entry, the distinction between prospective studies and retrospective studies is provided, followed by a discussion of the advantages and analysis of prospective studies and examples of influential studies.

### Prospective Versus Retrospective Studies

Prospective studies may be contrasted against retrospective studies, a term that is sometimes applied to case-control studies, in which subjects are selected based on the presence or absence of disease, and exposure history is then documented retrospectively. Prospective studies, however, may use data collected entirely or partially in the past (such as through analysis of existing records for exposure or both the exposure and outcome); in this case, the distinguishing feature remains that the exposure of interest existed, and was documented, prior to the onset of the condition. Across disciplines, there is some overlap in the use of the terms *prospective study, cohort study,* and *panel study.* All of these terms imply a prospective design (appropriate time sequence in the assessment of exposure and outcome).

In early publications, many authors used the term *prospective study* synonymously with the term *cohort study.* In essence, *retrospective* and *prospective* can be terms that are used in conjunction with study designs such as case-control and cohort. It is possible to have both prospective and retrospective case-control studies. A case-control study would be considered prospective if the exposure measurements were taken before the outcome is ascertained (often called a nested case-control study). Likewise, it is possible to have both a prospective and retrospective cohort study. A cohort

study can be classified as retrospective if the outcome has occurred before the exposure is ascertained. Cohort studies are often used as the foundation for related study designs such as case-control studies nested within cohort studies and case-cohort designs.

Confusion can exist given that both case-control and cohort designs can have both prospective and retrospective measurement elements. The prospective and retrospective distinction is often applied to the timing of subject identification and does not necessarily refer to the timing of outcome assessment relative to exposure assessment. For example, sometimes a cohort study is identified as a retrospective cohort study, indicating that it is a cohort study that involves the identification and follow-up of subjects, but subjects are identified after completing the follow-up period.

Experimental designs, such as a randomized control trial, would always be considered prospective because the study investigators assign subjects to various exposure groups and then follow these subjects forward in time to determine some outcome.

## Advantages

General advantages of prospective designs over case-control studies in terms of reduction of risk of bias include less opportunity for information error in documentation of the exposure (because this cannot be influenced directly by knowledge of the disease outcome). Potential sources of error in nonrandomized cohort studies include observer bias, respondent bias, and selection bias. Cohort studies may not include blinding to the exposure, and therefore observer and respondent bias may occur. If the exposure is known, study personnel may introduce a bias when assessing the study outcome. Likewise, if respondents are self-reporting, the outcome bias may be introduced.

## Analysis

Participants' may be selected purposively on the basis of being at risk of the outcome of interest and representing varying levels of the exposure of interest. Custom-designed prospective studies are often designed to select individuals at risk of the outcome within a feasible period of follow-up; for example, recruitment in middle age. Prospective studies have also been generated from representative samples (such as national-level health surveys) and combined with active or passive follow-up (such as vital statistics and hospitalization records). At the design or analysis stage, cohort studies typically require that participants be free of the condition. To ensure this, baseline screening may be used or a wash-out period may be used (e.g., excluding the first few years of mortality).

The outcome of cohort studies is often the incidence of first occurrence of the disease. Common forms of statistical analysis procedures are based on single time-to-event data structures, including most survival analyses. More sophisticated modern analytic approaches may take into account multiple occurrences of the outcome, time-dependent covariates and studies of competing risks, as well as sensitivity analyses for loss-to-follow-up.

## Historical Moments and Influential Prospective Studies

Historically, prospective studies have been conducted for more than 100 years. Richard Doll, in a 2001 article that examined the history of the method, reflected on his own work examining cigarette smoking and lung cancer. Doll noted, in response to his case-control study, that very few scientists at the time accepted the conclusion of the case-control study and that only a prospective study would be suitable. Hence, Doll and Austin Bradford Hill launched the British Doctors Study, a prospective cohort of 40,000 doctors examining their smoking habits with linkage to the vital statistics registry to obtain death records.

Further examples of influential prospective studies include the Framingham Heart Study, with more than 50 years of prospective follow-up, and the Nurses' Health Study, which started in 1976 and is tracking more than 120,000 registered nurses in the United States.

*Jason D. Pole and*
*Susan J. Bondy*

***See also*** Clinical Trial; Cohort Design; Longitudinal Design; Nested Factor Design; Retrospective Study

**Further Readings**

Belanger, C. F., Hennekens, C. H., Rosner, B., & Speizer, F. E. (1978). The nurses' health study. *American Journal of Nursing, 78*(6), 1039–1040.

Breslow, N. E., & Day, N. E. (1987). Statistical methods in cancer research. Volume II—The design and analysis of cohort studies. *IARC Scientific Publications, 82*.

Dawber, T. R., Meadors, G. F., & Moore, F. E. (1951). Epidemiological approaches to heart disease: The Framingham Study. *American Journal of Public Health & the Nation's Health, 41*(3), 279–281.

Doll, R. (2001). Cohort studies: History of the method. II. Retrospective cohort studies. *Sozial- und Praventivmedizin, 46*(3), 152–160.

Doll R. (2001). Cohort studies: History of the method. I. Prospective cohort studies. *Sozial- und Praventivmedizin, 46*(2), 75–86.

Rochon, P. A., Gurwitz, J. H., Sykora, K., Mamdani, M., Streiner, D. L., Garfinkel, S., Normand, S. L., & Anderson, G. M. (2005) Reader's guide to critical appraisal of cohort studies: 1. Role and design. *British Medical Journal, 330*(7496), 895–897.

# PROTOCOL

In research, *protocol* refers to the written procedures or guidelines that provide the blueprint for the research study, as well as good and ethical practices that should be observed when conducting research, such as good etiquette when dealing with participants, adherence to ethical principles and guidelines to protect participants, compliance with institutional review board requirements, not engaging in academic dishonesty, and so on. Good protocol is a critical component of high-quality research. A well-written research protocol is particularly important for obtaining institutional review board clearance for the research, as well as for seeking funding opportunities. In addition, the written research protocol serves as a manual to guide the entire research effort, and it can also be used as a monitoring and evaluation tool to monitor progress throughout the research and to evaluate success at the completion of the research. This entry describes the structure of the written proposal and discusses other types of protocols.

## Structure of the Written Research Protocol

The written research protocol is a detailed descriptive text of how the research will be conducted, and its length and complexity are affected by the nature and scope of the research. The principal investigator generally has some flexibility in determining the comprehensiveness and layout of a particular protocol; however, each research protocol must comply with the requirements (i.e., structure, context, format, length, etc.) for the institutional review board that will grant its approval. A well-written research protocol can contribute greatly to making a research effort high quality. Comprehensive information provides not only guidance and clarity on how to conduct each and every aspect of the research, but also advice on what should be done (e.g., whom to contact) if an unusual situation occurs that was unforeseen. The aforementioned structure and details are fundamental to good research protocols.

### *Cover Page and Title*

The cover page should include the full title of the study, as well as the version number, and version date (e.g., Version 1.0 dated May 1, 2010). The document should also indicate if the protocol is a "draft" or "final" document (e.g., Draft Protocol). The title should be short and concise, and also include key words such as the proposed research design, population to be investigated, and location of data collection. For example, a study on juvenile delinquency might be titled "Study of Juvenile Delinquency in Country A Using Convenience Sampling." These key words are important to facilitate classification/indexing of the project. A short title should also be specified for use throughout the document. The short title should be first abbreviated in the project summary. The cover page should also include the names, roles, and contact information for the primary authors/investigators/advisors, as well as any major sponsors. If this information is too much to place on the cover page, it should be placed on a separate page after the cover page.

### *Signature Page*

This page should include the signatures of the individuals listed on the cover page.

### *Contents Page*

This page details the various sections and appendixes contained in the protocol, along with corresponding page numbers.

### *Acronym/Abbreviation Page*

This page provides a list of all acronyms and/or abbreviations used in the protocol, along with a definition for each one.

### *Summary or Abstract*

The summary is the most important section of most documents/reports—the research protocol is no exception. It provides a succinct sketch (i.e., a snapshot) of the entire protocol. It is located at the front of the document but is generally prepared after the entire protocol is written. It summarizes the research objectives; presents the main research question or hypothesis; provides a brief description of the methods and procedures that will be used to conduct the research (e.g., research design, study population, place, timing of interventions); discusses the anticipated study outcomes; and provides the dates for important milestones as well as the estimated time frame and budgeted cost for the research.

### *Background*

The background describes existing knowledge and research in the area of study. It is supported by a comprehensive literature review of both published and unpublished work that should highlight any deficiencies or gaps in existing knowledge. The justification for the study, as well as a clear articulation of the research questions or hypotheses, and a detailed discussion of the study objectives (general and specific) are also generally included in this section. Some protocols, however, may use separate sections to discuss one or more of the aforementioned issues. For example, objectives may be discussed under a section titled "Specific Aims of the Study," research questions may be discussed under "Problem Statement," and so on. Alternatively, some protocols use a section called "Introduction" instead of "Background."

The justification should clearly outline the nature of the problem, its size, and its effect; the raison d'être for the study; the feasibility of the study; the potential benefits or outcomes that are expected to be derived from the study (categorized into participant, short-term/long-term, population, knowledge base, etc.); how the results might be used (e.g., national policy, organizational policy, community level, future research); how the study might help to correct the deficiency; how the study will add to the existing body of knowledge, and so on. The objectives should follow from the research questions or hypotheses—they should be smart, measurable, achievable, relevant, and time based (SMART).

### *Methodology*

The methodology provides a detailed description of the methods and procedures that will be used to conduct the study. It includes a description of the following, but may not necessarily be discussed under the separate subheads outlined below.

*1. Study Design.* This subsection describes the proposed study design (e.g., experimental, cohort, case-control) and provides justification for the choice. The selection of a particular study design is linked to the particular study objectives, as well as ethical considerations and availability of resources. Where an intervention is part of the research design, the protocol must discuss the intervention and note the following: frequency and intensity of intervention, venue, and entity or person in charge of the intervention.

*2. Sample Size, Selection, and Location.* The protocol should provide details on the sample size and the justification for the sample size (including chosen levels of significance and power), explain how the sample was selected (e.g., probability—random selection or nonprobability—nonrandom selection), explain how the locations and sites were selected, provide a clear listing of the inclusion and exclusion criteria for participant selection, and so on. Where informants and/or focus groups are being used, the selection criteria, number of persons participating, and so on will also need to be discussed.

*3. Unit of Analysis.* The unit of analysis—individual or group—must be discussed.

*4. Sampling Frame.* The protocol should discuss and justify the sampling frame selected for the study.

*5. Operationalization of the Variables.* All variables being investigated should be defined and operationalized.

*6. Measurement of Variables.* The protocol should specify how the variables will be measured.

*7. Incentives for Participation.* Any incentives that are being offered for participation, as well as the time and mode of delivery, should be outlined clearly.

*8. Data Collection Procedures.* Both primary and secondary data collection procedures (e.g., surveys—face-to-face/telephone/mail, focus groups, nonparticipant observation, content analysis) must be outlined and justified. In the case of secondary data sources, the source, content, and quality will also need to be discussed. Instruments being used, including interview guides, registration forms, and so on, must be appended to the protocol.

*9. Quality Control Procedures.* The protocol should clearly outline the strategies that will be used to safeguard data integrity, and thus by extension enhance overall (a) reliability (e.g., to ensure high interrater reliability, raters will be properly trained and monitored, and a code manual will be prepared to ensure that the coding process is consistent); (b) validity (e.g., face and content validity will be established by having an expert panel review the instrument); and (c) data quality (e.g., all interviewers will receive 12 hours of training that will include several demonstration exercises, the interview guide and the questionnaire will be discussed in-depth, and a pilot study will be conducted). The discussion on quality control may be incorporated under specific subheads.

*10. Data Management.* The procedures for data storage and data security should be discussed. This discussion should identify where the data will be stored or located, as well as the duration of retention (e.g., federal regulations in the United States require data retention for at least 3 years), who will have access to the data, the level of access, and so on. The procedures for data collection, coding, and validation should also be discussed.

*11. Project Management.* The procedures for managing the study should be outlined. Issues to be discussed will include staff requirements, staff training, supervision of personnel, responsibility matrix, the project schedule and work breakdown structure (generally prepared using Microsoft Project), and the project budget. If only an overview is provided of the project schedule and budget, then the detailed schedule and budget should be appended to the report.

*12. Data Analysis.* The software packages that will be used for data analysis should be described, as well as the types of analyses that will be performed (e.g., statistical, nonstatistical, analytical). In addition, other issues such as the handling of missing data, outliers, and so on should be discussed.

*13. Ethical Issues.* Ethical issues are central to research protocol. Any potential risks (e.g., physical, psychological, social, economic) to participants should be fully discussed. Measures to protect participants from possible risks or discomforts should also be fully disclosed. The manner in which informed consent is to be obtained should also be addressed. For example, in research involving certain groups (children, mentally challenged, elderly), informed consent must be obtained from a third party. In such cases, the protocol must explain how the third party will be contacted, and so on. Participants' rights to withdraw any time without risk of prejudice, penalty, or otherwise, as well as their rights to refuse to answer particular questions, should also be outlined clearly. The protocol should also outline who will provide debriefing and indemnity if participants are harmed through negligence or otherwise while participating in the study.

*14. Anonymity and Confidentiality.* Procedures for keeping the data anonymous and confidential should be outlined, as well as the procedure that will be used to assure participants that this will be done (e.g., official letter backed up by verbal assurance stating that all data will be aggregated and no individual-level data will be revealed). This is particularly important when sensitive data are involved. This discussion should be linked to data management. The data coding procedures for participant identification should also be discussed.

### *Limitations of the Study*

The limitations of the study should be clearly outlined so that the users of the information can make informed judgments in light of the limitations.

### *Dissemination or Publication of Results*

This section should discuss the stakeholder groups that will have access to the research findings and how the results will be communicated (e.g., written report, presentation, town hall meeting, journal article, news media). It should also discuss who will have publication rights.

### *References*

A list of all references quoted in the preparation of the protocol should be listed in sequential order.

### *Appendixes*

Research instruments, consent and assent forms, letter to participants assuring anonymity and confidentiality, interview guides, detailed schedules and budgets, and so on, should be appended to the protocol. The curriculum vitae of each principal and secondary investigator should also be appended. If advertising is done in order to recruit participants, all recruitment materials, such as press releases and radio, television, and newspaper advertisements, should also be appended to the protocol that is being submitted for institutional review board approval.

## Other Types of Protocol

In addition to the written protocol that is important for institutional review board purposes, written protocols for interview guides and schedules may also be prepared to guide the research effort in its quest for high-quality data. In addition, other types of protocol, such as appropriate behavior when dealing with participants (e.g., greeting, listening attentively to participants, keeping an appropriate distance from the participants, etc.) and dress protocol, are usually discussed in training sessions.

*Nadini Persaud*

*See also* Ethics in the Research Process; Informed Consent

### Further Readings

The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979, April). *The Belmont report: Ethical principles and guidelines for the protection of human subjects of research* [Online]. Retrieved May 1, 2009, from http://ohsr.od.nih.gov/guidelines/belmont.html

Pan American Health Organization. (n.d.). *Guide for writing a research protocol* [Online]. Retrieved May 1, 2009, from http://www.paho.org/English/DD/IKM/RC/Protocol-ENG.doc

Singh, S., Suganthi, P., Ahmed, J., & Chadha, V. K. (2005). *Fomulation of health research protocol: A step-by-step description* [Online]. Retrieved May 1, 2009, from http://medind.nic.in/nac/t05/i1/nact05i1p5.pdf

Western Michigan University. (2005, June). *Human subjects institutional review board: Application for project review* [Online]. Retrieved May 1, 2009, from http://www.wmich.edu/coe/fcs/cte/docs/hsirbproposal_vandermolen.doc

# "Psychometric Experiments"

"Psychometric Experiments" is an article written by Sir Francis Galton that described new methods for measuring human thought. He believed that if one wished to study the mind, there must be some replicable, verifiable means of measuring and quantifying its operation.

Beginning no later than 1876, Sir Francis Galton turned his extraordinarily prodigious mind to the development of a science of psychometry in Great Britain. At the same time that Wilhelm Wundt opened his experimental psychology laboratory in Germany, Galton conducted suggestive early research in psychometry. These two individuals' approaches were similar in some ways, but differed fundamentally in others. Both men were committed to the study of the human mind through introspection—Wundt by establishing rigorous experimental research methods and Galton by applying observational techniques to establish an understanding of the character of human thought. Wundt would likely agree with Galton that a discipline cannot attain the status of a

science until its phenomena have been measured reliably and validly.

Wundt influenced the early development of psychology to a far greater extent than Galton; Wundt's laboratory was associated with the University of Leipzig and attracted students internationally (e.g., J. M. Cattell, G. Stanley Hall, E. B. Titchener, and William James, to name a few), all of whom published prolifically for many decades. On the other hand, Galton, like his cousin Charles Darwin, was a wealthy individual and largely worked alone. His only disciple of considerable note was Karl Pearson. Consequently, psychological research during most of the 20th century was driven by the experimental and quasi-experimental designs fomulated by Wundt and his followers. It was only in the last quarter of the 20th century that Galton's (much-evolved) concepts of correlation and regression (e.g., multiple regression analysis, exploratory and confirmatory factor analysis, canonical correlation, and structural equation modeling) began to characterize much of psychological research.

Galton was particularly interested in the development of instruments with which to measure human mental attributes. For him, psychometry was the art of measuring and assigning numbers to the attributes of the mind (e.g., determining the response latency of individuals). British, European, and American psychologists adopted many of his measurement tools, and some remain in use today (e.g., the questionnaire, word association tests).

In his paper "Psychometric Experiments," Galton proposed to provide a new form of psychometric investigation in which he attempted to quantify the processes of the human mind. Two studies were reported in his paper. In the first "experiment," Galton walked down the Pall Mall a distance of some 450 yards, focusing intently on each successive object that caught his eye. When he had generated one or two thoughts directly associated with the object, he made a mental note of each idea, then moved on to the next object. Afterwards, he calculated that he viewed approximately 300 objects during his walk that evoked a wide array of memories formed throughout his life. He noted that he recalled previous incidents of which he was not consciously aware until after the recollection occurred. Galton repeated his 450-yard walk along the Pall Mall and discovered there were many repetitions of associations from his first walk. These results, although intriguing, were not systematically recorded, and thus they were not available for statistical analysis.

In an attempt to refine his previous study, Galton constructed a list of 75 printed words. He placed the list beside himself and positioned a book on top so that when he sat back in his chair, the words would be obscured. Subsequently, he would lean forward and only one word on the list would be visible at a time. Galton sat with his pen resting on a book in his right hand and a watch in his left. He then leaned back in his chair, allowed his mind to become blank, and leaned forward to read the word and simultaneously begin timing. Once one to four ideas culminated, Galton ceased timing, recorded the associations that had occurred, the word that had been displayed, and the time shown. Trials required a maximum period of about 4 seconds. He repeated this procedure on four separate occasions at 1-month intervals.

Galton conducted 300 separate trials during this study, which resulted in a total of 505 ideas during a time span of 660 seconds. Upon completion of his experiment, Galton expressed dismay at the fact that, of the 505 ideas suggested after the culmination of four sets of trials, only 289 of the ideas had occurred without repetition. He found he could place the origin of 124 of his associations in chronological order using the following categories: boyhood and youth, subsequent manhood, and quite recent events. The majority of his ideas resided in the first two categories (39% and 46%, respectively), with 15% in the third category. He concluded that earlier associations were better established than more recently formed associations.

Galton classified his ideas according to the nature of the word that was employed as the stimulus. He used three categories: (1) the imagined sound of words, such as one's name or verbal quotations; (2) sense imagery; and (3) histrionic types of actions that could be performed by oneself or others. He further classified the words presented into three groups: (1) words that represented a definite image, such as abbey, aborigines, and abyss; (2) words that spawned a histrionic association, such as abasement, abhorrence, and ablution; and (3) abstract words, such as afternoon, ability, and abnormal. Within the abbey category, 43% were sense images, 11% histrionic, and 46%

verbal. The abasement category consisted of 33% histrionic, 32% sense images, and 35% verbal. The afternoon category consisted of 53% verbal, 22% sense images, and 25% histrionic. At bottom, Galton found that abstract words were the most difficult about which to form associations; they took longer to generate and fewer associations were produced.

Although one may marvel at Galton's seemingly endless desire to quantify, measure, and investigate the human condition (his motto was "Whenever you can, count"), his psychometric studies lacked technical soundness. There were few controls, and when controls were mentioned, they were not clearly described. For instance, he claimed that he ensured the complete fairness of the experiment by instituting several small precautions, but failed to mention what they were. Despite this fact, Galton's paper provided an excellent qualitative narration of the process of free association, detailed his mental processes of verbal association and generalization, and noted that the human mind has a tendency to be recursive when dealing with stimuli. Galton proposed that ideas come to fruition as associations, such as when one views an object that evokes a fresh idea or causes one to revisit a previous notion. When associations have been formed with a particular stimulus, the mind has a tendency to return to the associations when the same stimulus is encountered again. Galton hypothesized that ideas were committed to long-term memory when current associations were revisited often and were fueled by an individual's sustained interest. The same memories become degraded when they are neglected across time.

Galton viewed the method of forming associations as automatic. He noted that this process would not normally be detected by an individual and would overlineely cross the threshold of his or her consciousness. He described the practice of bringing clandestine thoughts to light as laborious and mentally taxing. Galton described the unconscious as an essential component of the human mind. Although for some, the existence of unconscious mental processes seemed highly unlikely, Galton spoke of such processes as vital and acknowledged that if the human brain was limited to only the part of it that exists within one's consciousness, it would greatly interfere with one's ability to complete daily tasks. Contemporary researchers, such as Joseph LeDoux, have made similar statements and have described the unconscious as being responsible for a myriad of mental processes that encompass the majority of mental life (e.g., evaluating, believing, judging, feeling, and imagining).

Galton has been referred to as the father of differential psychology. He noted how two individuals shown the same stimulus could have very different reactions. In 1967, John Lacey coined the term *situational stereotypy* in reference to how individuals encode and react to the same stimuli differently. Galton spoke of the difficulty of attempting to compare two persons' mental processes and acknowledged that even if both individuals were presented with a verbal sentence at the same time, the first response to any given word would differ broadly with respect to each individual.

The introspective method that both Galton and Wundt used has been criticized for its use in systematic research. For example, Galton studied his own introspections in both naturalistic and laboratory situations, but he became progressively disillusioned with the method and eventually claimed that different brain components fail to work cooperatively. Indeed, he believed that conscious thought comprised but a small fraction of the brain's largely automatic activity. The concept of consciousness traditionally implies both an *awareness* of the process of thought and also a control of behavior through *intentions*. Galton clearly questioned both of these assumptions regarding the majority of human thought and behavior.

Although his research was clearly outstripped by the overwhelming number of publications coming from Wundt and his many disciples, Galton's work in instrument development and statistical analysis was second to none. Over the course of his life, he published more than 58,000 pages of research findings. Considering his achievements in geographical exploration and mapping, meteorology, fingerprint analysis, causes of snoring, the intelligence of earthworms, and many more endeavors, it is difficult to imagine how he found time for his work in human psychology. Had he devoted his entire life to psychological studies, one can only speculate what additional discoveries he might have accomplished.

*Ronald C. Eaves, Suzanne Woods-Groves, and Thomas O. Williams, Jr.*

*See also* Natural Experiments; Psychometrics; Thought Experiments

### Further Readings

Crovitz, H. F. (1970). *Galton's walk: Methods for the analysis of thinking, intelligence, and creativity.* New York: Harper & Row.

Diamond, S. (1977). Francis Galton and American psychology. *Annals of the New York Academy of Sciences, 271,* 47.

Editor. (1879). Mr. F. Galton on generic images and automatic representation. *Mind, 4,* 551–557.

Evans, J. St. B. T. (1983). Introduction. In J. St. B. T. Evans (Ed.), *Thinking and reasoning: Psychological approaches* (pp. 1–15). London: Routledge and Kegan Paul.

Galton, F. (1869). *Hereditary genius.* London: Macmillan.

Galton, F. (1879). Psychometric experiments. *Brain: A Journal of Neurology, II,* 149–162. Retrieved May 2, 2008, from http://www.galton.org/psychologit/index.html

Galton, F. (1879). *Psychometric facts.* Nineteenth Century, V(25), 425–433.

Lacey, J. I. (1967). Somatic response patterning and stress: Some revision of activation theory. In M. G. H. Coles, J. R. Jennings, & J. A. Stern (Ed.), *Psychophysiological perspectives festschrift for Beatrice and John Lacey* (pp. 42–70). New York: Van Nostrand Reinhold.

Merton, R. K., Sills, D. L., & Stigler, S. M. (1984). The Kelvin Dictum and social science: An excursion into the history of an idea. *Journal of the History of the Behavioral Sciences, 20,* 319–331.

Pearson, K. (1914). *The life, letters and labours of Francis Galton* (Vols. I, II, IIIA, IIIB). Cambridge, UK: Cambridge University Press.

## Psychometrics

Psychometrics refers to the measurement of abilities, traits, and attitudes with questionnaires and tests. While such techniques as factor analysis and structural equation modeling can be thought to fall under the umbrella of psychometrics, this article focuses on item response theory (IRT), which studies the relationship between traits of individuals and their responses to items on a test. Although the roots of IRT are in educational testing, it has applications in many other disciplines. Much of the pioneering work in the field is attributed to Frederick Lord.

The trait studied in IRT is most often a latent trait, one that cannot be measured directly. IRT describes how a person's observable behavior, or his or her performance on a test, is related to this latent trait or ability. Characteristics of each item on the test determine the likelihood of a particular response at a given level of ability. The purpose of IRT modeling is usually to estimate respondents' abilities, although the item characteristics may also be of interest.

Test items need to discriminate well in order to inform researchers about respondents' abilities. That is, people with lower values of the trait should tend to respond one way to an item while people with higher values of the trait respond differently. The relationship between values of the trait under consideration and the probability of a particular response is modeled with the item characteristic curve (ICC), sometimes called an item response function. The most basic IRT models are presented first, and then more sophisticated models are discussed.

### Logistic and Ogive Models

Logistic and ogive curves are natural choices for modeling ICCs when items have only two response options (or are coded to be binary) because they are monotonically increasing in the latent trait and they allow the probability of a correct response, conditional on the trait, to vary from 0 to 1. Ogive curves are defined by the cumulative normal distribution and logistic curves by the cumulative logistic distribution.

Before stating the models, the notation used to refer to the test is given. $\theta$ is the ability parameter or level of the trait, and the true ability of individual $i$ is denoted $\theta_i$ . Items are indexed with the letter $j$, and $Y_{ij}$ is the response of individual $i$ to item $j$. The number 1 denotes a correct answer and 0 an incorrect answer. $P_j(\theta)$ represents the probability that an individual with ability $\theta$ answers item $j$ correctly; it is the height of the ICC curve for item $j$ at $\theta$. For notational convenience in working with logistic curves, define $\psi(\theta, a, b) = \{1 + e^{-a(\theta - b)}\}^{-1}$. This is the height of the cumulative distribution function (c.d.f.) of the logistic distribution with location parameter $b$ and scale parameter $a$ at $\theta$.

Also, define $\Phi(\theta, a, b)$ to be the height of the normal cdf with mean $b$ and standard deviation $a$ at $\theta$.

The Rasch model, or one-parameter logistic model, is the most basic of the logistic models for IRT. The single-item parameter $b_j$ is a location parameter that is interpreted as the difficulty of item $j$. Under the Rasch model,

$$P_j(\theta) = \frac{1}{1 + e^{-(\theta - b_j)}} = \psi(\theta, 1, b_j).$$

Thus, the limit as $\theta$ approaches negative infinity of $P_j(\theta)$ is 0 and the limit as $\theta$ approaches infinity is 1. The center of the curve, $b_j$, is the point at which $P_j(\theta) = 1/2$.

The two-parameter logistic model (2PL) adds a scale or discrimination parameter $a_j$ to the Rasch model. Under the 2PL model,

$$P_j(\theta) = \frac{1}{1 + e^{-a_j(\theta - b_j)}} = \psi(\theta, a_j, b_j).$$

To maintain the increasing monotonicity of the ICC, $a_j$ must be positive. The limits of this curve are the same as for the Rasch model, and $P_j(b_j)$ is also 1/2. $a_j$ is the slope of the ICC when $\theta = \mathrm{b}_j$. In the Rasch model, $a_j$ is fixed at 1. Larger values of $a_j$ indicate better discrimination for item $j$. Items with better discrimination provide more information about individuals' abilities.

The three-parameter logistic model (3PL) adds a parameter $c_j$ to account for the probability of guessing the correct answer. The ICC for the 3PL model is specified by

$$P_j(\theta) = c_j + \frac{1 - c_j}{1 + e^{-a_j(\theta - b_j)}}$$

$$= c_j + (1 - c_j)\psi(\theta, a_j, b_j),$$

with $0 \leq c_j \leq 1$. The lower asymptote of this curve is $c_j$ rather than 0, and the point at which $P_j(\theta) = 1/2$ is shifted to the left of $b_j$.

In the two-parameter normal ogive model,

$$P_j(\theta) = \Phi(\theta, a_j, b_j).$$

$b_j$ is again the point at which $P_j(\theta) = 1/2$, and $a_j$ determines the slope of the curve. The shapes of the normal ogive and logistic curves are very similar to one another. Early work in the field began with the normal ogive, but logistic models are now more popular for ease of computation. Sometimes, a scaling constant is used with a logistic curve to match it as closely as possible to the ogive curve.

The next subsection discusses methods for fitting a basic IRT model in which individuals are assumed to respond independently of one another, items are assumed to be independent of one another, and individuals and items are independent. It is also assumed that the test is unidimensional, measuring a single trait. In this case, the likelihood of observing the data matrix $Y$ is

$$\prod_{i=1}^{N} \prod_{j=1}^{M} P_j(\theta_i).$$

Here, $N$ denotes the number of respondents and $M$ the number of items.

## Model Fitting and Software

Analyzing a test with the above IRT model involves estimation of both the parameters of the ICCs and the ability levels of the test takers. There are many parameter estimation techniques, a few of which are mentioned here.

Joint maximum likelihood estimation (MLE), proposed by Alan Birnbaum, alternates between estimating abilities and item parameters until these estimates meet a convergence criterion. Initial ability estimates based on the raw test score are used and the Newton-Raphson technique is used to find the MLEs. Fixing the location and scale parameters of the ability distribution ensures identifiability of the curve parameters. The computer program LOGIST uses this method. The algorithm is not well-suited for fitting the three-parameter logistic model, however, as the estimates often fail to converge.

Marginal maximum likelihood estimation, proposed by R. Darrell Bock and Marcus Lieberman and refomulated by Bock and Murray Aitkin, does not involve iterating through item and ability estimates. It is assumed that the examinees represent a random sample from a population where abilities follow a known distribution, most commonly the standard normal distribution. This allows the analyst to integrate the likelihood over the ability distribution and estimate item parameters in the marginal distribution via the expectation-maximization algorithm. The

computer program BILOG implements this procedure.

A problem with maximum likelihood methods is that item parameters cannot be estimated for items that all respondents answer correctly or all respondents answer incorrectly. Similarly, abilities cannot be estimated for individuals who respond correctly to all items or incorrectly to all items. Such individuals and items are discarded before maximum likelihood estimation is carried out. A Bayesian approach to modeling IRT resolves this problem, retaining the information about abilities contained in these items and also allowing the analyst to incorporate prior information about model parameters. Prior distributions are placed on all model parameters, and then Bayes's rule is used to update the information about model parameters using the data. This results in posterior distributions for model parameters that are used for inference. A hierarchical Bayesian version of the two-parameter logistic model may be defined as follows:

$$\begin{aligned} Y_{ij}|\theta_i, a_j, b_j &\sim Bernoulli\big(\psi(\theta_i, a_j, b_j)\big) \\ \theta_i|\mu_\theta, \sigma_\theta^2 &\sim Normal(\mu_\theta, \sigma_\theta^2) \\ b_j|\mu_b, \sigma_b^2 &\sim Normal(\mu_b, \sigma_b^2) \\ a_j|\mu_a, \sigma_a &\sim \log Normal(\mu_a, \sigma_a). \end{aligned}$$

The discrimination parameter $a_j$ is sometimes given a gamma or truncated normal prior distribution. James H. Albert proposed using the Gibbs Sampler to fit a model similar to this one and to compute posterior means and standard deviations for item and ability parameters. A Metropolis-Hastings within Gibbs sampling strategy can also be used for this problem, and both can be extended for more complex models. WinBUGS software can be used to fit this model.

## Test Information

Item and test information functions indicate how well ability is estimated over the ability scale. The Fisher information for $\theta$ is the reciprocal of the variance of the MLE for $\theta$ when item parameters are known. The test information function is given by

$$I(\theta) = \sum_{j=1}^{M} \frac{[P'_j(\theta)]^2}{P_j(\theta)[1 - P_j(\theta)]}.$$

Information depends only on item characteristics, not individuals' responses. It is highest for abilities in regions where the ICCs are steepest and adding items to a test increases the information. Administrators of large testing operations typically have a test bank of items whose characteristics are known and thus can design a test with the test information function in mind. If the aim of the test is to measure abilities accurately across a wide range of abilities, a mixture of difficulties is called for, whereas a test to determine if respondents meet a particular threshold ability should have many questions with difficulty levels near the threshold. Computerized adaptive testing involves estimating respondents' abilities as they take the test and choosing the next items from an item bank in such a way as to maximize the information. This generally leads to shorter tests and more accurate estimates of ability for individuals in the tails of the ability scale.

With a variety of IRT models available for explaining test responses, the questions of model choice and assessment arise. Models produce estimates of both individual abilities and the ICCs. Models with better fit should estimate both of these sets of parameters well. The fit of the ICC is often considered on an item-by-item basis. It is important to note, however, that item fit statistics are dependent upon the accuracy of ability estimates. For short tests (fewer than 15 items), ability cannot be estimated well, and ICCs, even when correctly modeled, may not appear to fit well.

A common approach for measuring the fit of a model with a categorical response is to compare observed and expected proportions of correct responses for individuals who fall in intervals of the predicted value of the latent trait. Ronald K. Hambleton and H. Swaminathan provided a list of methods for checking the model predictions. They first recommend visual inspection of "residuals." Individuals are placed in groups based on sorted values of ability estimates. The fitted curve is plotted along with the group average estimated ability versus group proportion correct. Along these same lines, a chi-square statistic may be calculated by forming 10 groups based on sorted ability and forming the $2 \times 10$ table of observed and expected frequencies where the rows correspond to correct and incorrect responses. Let $\hat{P}_{ij}$ represent the probability that respondent $i$ answers item $j$ correctly

according to the model, and let $\overline{\hat{P}}_{kj}$ be the average of these probabilities for individuals in ability group $k$. Also, let $O_{kj}$ be the observed number of correct responses by individuals in group $k$, and let $n_k$ be the number of individuals in group $k$. The statistic for item $j$ is given by

$$Q_j = \sum_{k=1}^{10} \frac{\left(O_{kj} - n_k \hat{P}_{kj}\right)^2}{n_k \overline{\hat{P}}_{kj}\left(1 - \overline{\hat{P}}_{kj}\right)}.$$

A high value of the statistic indicates poor fit. These chi-square statistics can be difficult to use in practice as items with low or high difficulty often end up with low cell counts. They also may not have very high power for detecting misfit as deviations from the curve get averaged over a range of abilities. Cross-validation is a good tool for assessing model fit, but can be too computationally intensive. A few other model fit techniques have been proposed, but a universally satisfactory statistic or method has yet to be established.

IRT allows one to evaluate the abilities of individuals who have taken different tests measuring the same latent traits by directly comparing their ability estimates. The fact that the model incorporates item characteristics makes possible this easy comparison of tests that may have different difficulty levels. This is in contrast to classical test theory, an earlier model for testing, which decomposes the observed score on a test into a "true" score and an error term. True scores are test-specific, and thus comparing abilities of individuals who have taken different tests may not be possible. Test equating is a large subfield of psychometrics. The literature gives recommendations for designing tests for equating scores of similar populations at different times (horizontal equating) and for equating tests given to populations of differing abilities such as different grade levels in school (vertical equating).

## Extended Models

IRT is a rich and actively researched field. In this section, a few extensions of the basic IRT model are discussed.

One assumption of IRT mentioned earlier is that of unidimensionality. Factor analysis techniques can be used to assess the validity of this assumption. When unidimensionality is not satisfied, as may be the case in a mathematics exam that tests both computational ability and critical thinking skills, the incorrect model specification leads to bias. A few multidimensional IRT models have been proposed thus far, including the Multidimensional 2PL Model and the Multidimensional Random Coefficients Multinomial Logit Model.

One way that independence of respondents may be violated is with differential item functioning. Differential item functioning refers to differences in the probability of a correct response to an item for groups of respondents having the same standing on the latent ability measured by a test. Checking for and understanding the effects of differential item functioning are crucial for maintaining the fairness of tests.

When groups of items are dependent on one another, as is the case when a set of questions refers to the same reading passage, a testlet model may be employed. Several Bayesian testlet models that account for item dependence have been proposed.

Many questionnaires use scaled response, and test items may also have a scaled or ordinal response. Fumiko Samejima's graded response model handles this type of data. The idea behind graded response is that a respondent's hypothetical "true" response falls along a continuous scale, and the scale categories offered as responses encompass a range of responses on the continuum. The probability of giving a particular response is calculated as the area under a normal curve in the region of the range of the response. The mean and variance of the normal curve are conditional on the respondent's trait level. Graded item response models can be fit with the programs MULTILOG and PARSCALE. Other models for polytomous items include the partial credit model, generalized partial credit model, and the rating scale model.

Models also exist for nominally scored data. In standardized tests that use a multiple-choice format, the responses are typically scored as correct or incorrect and analyzed with a model for binary data. Information about a respondent's ability can be lost in this recoding of data as ability level may influence which wrong response is chosen. Bock's nominal model makes use of the multivariate logistic distribution to provide item response functions

for the nominal responses. MULTILOG also fits this model.

The logistic and normal ogive parametric models that are the base for all models described so far assume a specific shape for the ICC. If the model is incorrect, ability estimates are biased. Nonparametric approaches offer more flexible models for the ICC that may provide a better fit than the parametric alternative. Kernel smoothing, or local averaging, is one approach for nonparametric estimation of ICCs. It is implemented in J. O. Ramsay's computer program Testgraf. Raw scores are used for initial estimates of ability, and the height of the ICC at a point is a weighted average of responses of examinees with abilities near this value. After the ICCs are computed, new MLEs for ability are estimated. This method may result in nonmonotonic curves if restrictions are not placed on the kernel smoother. Bayesian semiparametric and nonparametric methods for modeling monotonic ICCs make use of Dirichlet process priors.

IRT has provided the community of psychometric practitioners with a versatile tool for modeling tests and abilities. Advances in the field are bringing more accurate models, which in turn give more accurate ability estimates. This is an important achievement with the prevalence of high-stakes testing in today's society.

*Kristin Duncan*

***See also*** Classical Test Theory; Differential Item Functioning; Item Response Theory; Latent Variable; Structural Equation Modeling

**Further Readings**

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics, 17,* 251–269.

Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques*. New York: Marcel Dekker.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating*. New York: Springer.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

# Purpose Statement

A purpose statement is a declarative statement that summarizes a research project's main goal or goals. A purpose statement provides some guidance in establishing a research question and serves as an introduction to the resultant paper or dissertation chapter.

## Developing the Purpose Statement

At the beginning of a research project, it is helpful for the researcher to use a declarative sentence to state the main goal or goals of the project in specific terms. Statements that begin with the phrase "I wish to learn . . ." or "I plan to examine . . ." can be helpful insofar as they can move the topic's abstract notions to a concrete research question, which is the springboard for the resultant research design. Furthermore, a purpose statement can ground the researcher, providing a point of reference to which the researcher may return, particularly as the study increases in complexity. However, this does not imply that the purpose statement is final, because the researcher may revise the statement as needed. If external factors such as unavailability of relevant data force the researcher to make substantial changes to his or her research design, he or she may want to update the purpose statement to reflect those changes.

## Using the Purpose Statement

In addition to serving as a catalyst for the underlying research project, a purpose statement can be worked into subsequent papers or dissertation chapters derived from the project. Always near the end of the introduction, a purpose statement states the paper's intent, scope, and direction. Specifically, it provides for an abbreviated preview of the paper's main topic, while avoiding a discussion of the author's specific conclusions.

In research papers, purpose statements often start with phrases such as "This paper examines . . . ,"

"The main purpose of this study is to …" or "The aim of this article is to …" Purpose statements should be specific and precise, and should avoid vague, ambiguous, or confusing language. This ensures that there is no doubt in the reader's mind as to the research project's intended direction.

### Introductions: Purpose Statements Versus Thesis Statements

A purpose statement also serves as the foundation for a thesis statement, which provides assertions about the topic at hand and summarizes the author's conclusions. Unlike a purpose statement, a thesis statement provides a cursory answer to the question and is developed after the researcher has gathered evidence, which is presented in the body of the research paper.

The decision to use a thesis statement in the introduction is determined by the underlying norms of the specific discipline, as well as the author's preferences. In some cases, the author may simply state the paper's intended purpose at the outset, delaying the discussion of any results until the end of the paper. At the very least, a research paper introduction should provide a discussion of the research question and some information about how the author intends to explore the question, even if the answers are not presented until the conclusion.

### Examples of Thesis and Purpose Statements

*Ineffective purpose statement #1:* "This paper examines the impact of elites upon elections." It is unclear about what types of elites, what types of elections, or even which potential electoral effects the researcher intends to examine.

*Effective purpose statement #1:* "This paper examines the extent to which public endorsements by political elites shape electoral participation, particularly in proposition elections where the traditional information shortcut of partisanship is absent."

*Thesis statement #1:* "Elite cues help increase electoral participation because they provide information shortcuts to potential voters who may not be fully informed about the details of a given electoral contest."

*Ineffective purpose statement #2:* "This paper examines changes to election laws." In addition to being nonspecific as to what types of election laws are being examined, it is unclear as to whether the author is examining the sources of those changes, or the potential impact of those changes.

*Effective purpose statement #2:* "This paper examines the potential impact of the straight-ticket voting option upon electoral down-ballot outcomes."

*Thesis statement #2:* "Although Illinois Republicans appeared to eliminate the straight-ticket option in 1997 for partisan reasons, evidence suggests that Democrats might have actually benefitted from this ballot format change."

*Michael A. Lewkowicz*

***See also*** Dissertation; Research; Research Question

**Further Readings**

Popper, K. (1959). *The logic of scientific discovery.* New York: Basic Books.

## *p* Value

*p* values are calculated as a part of hypothesis testing, and *p* values indicate the probability of obtaining the difference observed in a random sample or a more extreme one in a population where the null hypothesis is true. Because of the widespread use of hypothesis testing, *p* values are a part of virtually all quantitative research reports. This entry explains *p* values in connection to other aspects of hypothesis testing and provides some cautions concerning the use of *p* values.

### Hypothesis Testing

Hypothesis testing is one of the main methods used for statistical inference. In hypothesis testing, researchers set up a hypothesis about a population parameter(s) and, based on data from a random sample drawn from this population, test its

tenability. The tested hypothesis is called the null hypothesis. Null hypotheses are represented by $H_0$. Null hypotheses may specify a value or a range of values for population parameter(s), differences or relationships between population parameters, or an effect in the population. For example, $H_0$ can state that a population mean is 100 ($H_0 : \mu = 100$) or that the difference between two populations' means is greater than 50 ($H_0 : \mu_1 - \mu_2 \geq 50$). In the most common application of hypothesis testing, $H_0$ of no association, no difference, or no effect is used, and this type of $H_0$ is often referred to as a nil null hypothesis. An example of a nil null hypothesis might be that there is no difference between two population means ($H_0 : \mu_1 = \mu_2$).

The $H_0$ is tested against a statistical alternative hypothesis. If $H_0$ is rejected, the alternative hypothesis holds true for the population. Alternative hypotheses are commonly symbolized as $H_a$, although some use other symbols, such as $H_1$. Generally, $H_0$ is tested using data from a sample(s), however, both $H_0$ and $H_a$ always refer to population parameters. For the $H_0$ that there is no difference between two population means ($H_0 : \mu_1 = \mu_2$), an $H_a$ would be that these population means are not the same ($H_a : \mu_1 \neq \mu_2$). A nondirectional $H_a$ about parameters is called a *two-sided* or *two-tailed* $H_a$. For example, $H_a : \mu_1 \neq \mu_2$ denotes a difference between two population means but does not indicate which of these means is larger or smaller. On the other hand, a one-sided or one-tailed $H_a$ specifies the direction of the relationship. For the $H_0 : \mu_1 = \mu_2$, a one-sided $H_a$ would be $H_a : \mu_1 < \mu_2$ or $H_a : \mu_1 > \mu_2$.

Often, $H_a$ states researchers' expectations or assumptions about the study's outcome and is commonly referred to as the research hypothesis. In many cases, researchers would like to reject $H_0$ and hold $H_a$ tenable for the population. However, this is not always the case. For example, when testing the effectiveness of a cheap drug compared to a more expensive drug, $H_0$ can be that the cheap drug's effectiveness is the same as the expensive drug's. In this situation, researchers probably would not want to reject the $H_0$.

After establishing $H_0$ and $H_a$, data are collected from a sample and then used to test the tenability of the hypotheses about population parameters. The hypothesis that is tested is always $H_0$, not $H_a$. The hypothesis testing process is a form of indirect proof—proof by contradiction. The testing process starts with the assumption that $H_0$ is true for the population. If the magnitude of the difference between the obtained statistic (e.g., $\overline{X}$ or *SD*) and the population parameter as stated in $H_0$ is highly unlikely to be observed in a sample belonging to a population where $H_0$ is true, then $H_0$ is *rejected* in favor of $H_a$. If the observed difference is not sufficiently unlikely, then $H_0$ is considered to be tenable for the population, and researchers *fail to reject* $H_0$. Notice that researchers either *reject* or *fail to reject* $H_0$; $H_0$ is never accepted because it is never proved, but rather the evidence may be insufficient to disprove it.

In hypothesis testing, researchers use a sample to infer about the population(s). However, every sample may or may not reflect the whole population. Therefore, researchers who have to rely on the sample statistic to reach a conclusion regarding the hypothesis about the population(s) face the probability of two types of errors:

> *Type I error* occurs when $H_0$ is rejected in favor of $H_a$ based on the sample statistics, when in fact, the $H_0$ is true in the population(s).
>
> *Type II error* occurs when $H_0$ is not rejected based on the sample statistics, when in fact, the $H_0$ is false in the population(s).

As seen in Table 1, when researchers make a decision about rejecting or failing to reject $H_0$, they face the probability of either a Type I or a Type II error, but not both. If $H_0$ is rejected, only Type I error probability should be considered, and if $H_0$ is not rejected, then only Type II error probability remains. Although it is desirable to eliminate both types of errors, this is not possible. However, hypothesis testing allows for controlling Type I error rate.

At the onset of a study, researchers choose a probability level for Type I error that they think is acceptable when consequences of an error are

**Table 1** Type I and Type II Errors

| | *State of the $H_0$ in the Population* | |
|---|---|---|
| *Statistical Decision* | *True $H_0$* | *False $H_0$* |
| Reject | Type I error | Correct |
| Fail to Reject | Correct | Type II error |

considered. Typically, if other variables remain constant, when Type I error probability decreases, Type II error probability increases, and vice versa. Therefore, when researchers specify a level for Type I error rate, they need to consider the consequences of both types of errors. For example, in a drug trial, incorrectly deciding that a new drug with serious side effects is more effective for improving patient survivability than a currently available drug with mild side effects—thereby committing a Type I error—can lead to exposing patients to serious side effects without any other health benefits. In such a case, researchers are likely to set a low probability level for Type I error. On the other hand, in a study where a promising new instructional model is evaluated against traditional instruction, incorrectly deciding that the new model is no more effective than the traditional model—thereby committing a Type II error—may result in abandoning research in a favorable area. In such a situation, which does not involve dire consequences, researchers may opt for a greater probability of Type I error.

## Level of Significance

The probability of Type I error in hypothesis testing is controlled by the chosen level of significance and is commonly represented by alpha ($\alpha$) or $p_{\text{critical}}$. Alpha can range between 0 and 1 like any probability value and is chosen before data collection. The most commonly used alpha values are .05 and .01, although researchers can use other alpha levels as they consider Type I and Type II error consequences. An alpha of .05 means that the probability of making a Type I error (rejecting the $H_0$ when it is actually true) is 5%. Accordingly, an alpha of .01 means that there is a 1% probability of a Type I error. Alpha is used as the criterion for rejecting $H_0$.

Probability values in hypothesis testing are obtained from sampling distributions. For the statistic of interest (e.g., $\overline{X}$ or *SD*), a sampling distribution is derived mathematically such that the distribution represents statistics from an infinite number of same-size random samples as the actual sample from a population where $H_0$ is true. The probability of any range of statistics on a random sample from a population where $H_0$ is true is equal to the proportion of the area covered by that range in the sampling distribution. For example, for a sample size of 30, if the probability of obtaining a mean of $100 < \overline{X} < 150$ is 5%, then in the sampling distribution, which consists of means for every sample at size 30, the area covered by the range of 100 to 150 is 5% of the whole area. The level of significance is the proportion of area in the sampling distribution that is equal to the probability of Type I error.

## *p* Value

After alpha is set, researchers collect data on a sample drawn at random from a population(s) for which they will make inferences. It is important to remember that $H_0$ is set up before data collection. Using data from this sample, the statistic of interest is calculated. Decisions about rejecting $H_0$ are based upon the likelihood of the observed difference between the sample statistic and the hypothesized population parameter in a population where $H_0$ is true. Therefore, the next step is finding the probability of getting the observed difference or a more extreme one on a random sample from a population where $H_0$ is correct. This probability is called *p* value or $p_{\text{calculated}}$. If the *p* value is less than the alpha level, then $H_0$ is rejected. The decision to reject $H_0$ is called a *statistically significant* result. If the *p* value is greater than the alpha level, then $H_0$ is not rejected, and this decision is referred to as a *statistically nonsignificant* result.

The *p* value, like alpha and all other probabilities, ranges between 0 and 1 and is calculated using sampling distributions. Before explaining *p* value calculations, it is important to understand test statistics.

## Test Statistic

After the statistic of interest is calculated using data from a sample, researchers calculate how much the observed statistic in the sample differs from the population parameter stated in $H_0$. In any case, researchers would not expect to obtain the same value hypothesized for the population parameter from a random sample because sample statistics will vary around the population parameters. A test statistic quantifies the difference between the sample statistic and the hypothesized population parameter and is a standard score. If

the probability of obtaining the observed test statistic or a more extreme one in a population where $H_0$ is true (i.e., $p$ value) is less than the level of significance, then $H_0$ is rejected. If the $p$ value is larger than the alpha, $H_0$ is not rejected.

Test statistics are calculated using sampling distributions of the sample statistics. Test statistics differ based on the sampling distributions used. The most common test statistics are $z$, $t$, $F$, and $\chi^2$. Because there is not a unique test statistic used for every $H_0$ test, rather than the test statistic itself, a probability statement regarding the test statistic (i.e., $p$ value) is used to make the decision about $H_0$. The $p$ value provides a more global language to determine if the statistic differs substantively enough from the hypothesized population parameter to reject $H_0$.

To obtain the probability of any test statistics, that particular test statistic's sampling distribution is used. The test statistic sampling distribution is different from the aforementioned sampling distribution, which was the sampling distribution of the statistic of interest. Sampling distributions of test statistics are commonly available in statistics books, on the Internet, or as a function in software. As an example, if the calculated test statistic is a $t$-statistic, the sampling distribution for the $t$-statistic is used to calculate the probability of the observed $t$-statistic.

## Example

A heuristic example is used to illustrate $p$-value calculation. Although the sampling distribution and the test statistic will differ, steps used in this example for $p$ value calculation can be generalized to other hypothesis testing situations. This example, which compares male and female students' mathematics anxiety levels, is provided to illuminate processes underlying $p$ value calculations, so some hypothesis testing details will not be explained or elaborated. Before data collection, as the first step, $H_0$ and $H_a$ were stated as follows:

$H_0$ : The population means are the same.

$H_a$ : The population means are not the same.

Second, $\alpha = .05$, in other words, it is decided that 5% probability is an acceptable probability of committing a Type I error. Then, researchers collect data on randomly sampled 40 female and 40 male students. In these samples, the mean mathematics anxiety level for females and males was 200 and 180, respectively. Although the means are not the same in the samples, it cannot be directly concluded that this would also be the case in the population. Because of sampling fluctuation, statistics from random samples will vary around population parameters. Therefore, the probability of obtaining such a difference or a more extreme one in a population where $H_0$ is true will be obtained.

The third step involves test statistic computation. As mentioned before, test statistics are calculated using sampling distributions of the statistics of interest that are derived mathematically assuming $H_0$ is true in the population. In this example, the statistic of interest is the mean difference; therefore, the test statistic will be computed using the sampling distribution of mean differences derived assuming the mean difference in the populations is 0. A general fomula that can be used in many test statistic calculations is

$$\text{Test statistic} = \frac{\text{Sample statistic} - \text{Hypothesized parameter}}{\text{Standard error of the statistics}}.$$

For this example, assume the standard error of the mean differences is 9 and the test statistic to be computed is a $t$-statistic.

$$t = \frac{(200 - 180) - 0}{9} = 2.2$$

Thus, a mean difference of 20 in the sample is 2.2 standard deviations away from the hypothesized population mean difference, which is 0.

The last step is finding the $p$ value, which is the probability of obtaining a $t$-statistic of 2.2 or greater in a population where $H_0$ is true. Because in this example a two-sided $H_a$ is used, the probability of both $t \leq 2.2$ and $t \geq 2.2$ are needed. To obtain $p$ values, a $t$-statistic sampling distribution is used. Using a commonly available $t$-statistic sampling distribution, the probability of $t \leq 2.2$ is approximately .017. Because $t$-distributions are symmetrical, the probability of $t \geq 2.2$ will be approximately .017 also. Therefore, the probability of $t \leq 2.2$ or $t \geq 2.2$ (i.e., $p$ value) is around

.017 + .017 = .034, which is smaller than the alpha value of .05. Because the *p* value is smaller than the alpha, the decision about $H_0$ is to reject. If alpha was set at .01 instead of .05, then the decision would be to not reject because the *p* value is larger than .01.

## Careful Use of *p* Values

Quantitative results are often misinterpreted because the purpose of hypothesis testing in an applied setting can be complex. Partial understandings of hypothesis testing can result in systemic issues that can negatively affect the quality of research reports. For quality research reports, it is important to understand hypothesis testing and *p* values and use them thoughtfully. Therefore, a few comments are provided for the careful use of *p* values.

1. *p* values are substantively affected by sample sizes. When a large sample size is obtained, even very small differences or effects become statistically significant. Therefore, it is fundamental that researchers provide an estimate of practical significance (i.e., effect size) when reporting *p* values.

For illustration, think of two experiments each comparing two groups' means. Experiment 1 consists of two groups (Groups 1 and 2) with 10 members each. The means for Group 1 and Group 2 are 55 and 56, respectively. Experiment 2 also consists of two groups (Groups 3 and 4), but this time with 110 members each. The mean for Group 3 is the same as Group 1, 55, and the mean for Group 4 is the same as for Group 2, 56 (see Table 2). The question in both experiments is, "Are the two means statistically significantly different with $\alpha = .05$?" In Experiment 1, the $p_{\text{calculated}} = .571$, so the means are *not* statistically significantly different, but in Experiment 2, the $p_{\text{calculated}} = .043$, so the means are statistically significantly different, although the difference in the means in both experiments is exactly 1 point (see Table 2). This leads to the question, "Which decision is correct?" To examine this phenomenon, one should consider how important the 1-point difference is in both experiments. For this, effect size estimates need to be calculated. This exemplifies that using a large-enough sample size will result in statistically significant results even when effects are inappreciably small.

**Table 2** Descriptive Statistics

| | *Group* | *Mean* | *N* | *SD* |
|---|---|---|---|---|
| Experiment 1 | 1 | 55.0 | 10 | 5.3 |
| | 2 | 56.0 | 10 | 6.7 |
| Experiment 2 | 3 | 55.0 | 110 | 5.3 |
| | 4 | 56.0 | 110 | 6.7 |

2. Another implication of *p* value's sensitivity to sample size is that sometimes, statistical significance cannot be obtained because of the small sample size, although the effect might be present in the population. Therefore, as researchers plan their studies, they are encouraged to do a power analysis to determine the appropriate sample size to improve the probability of correctly rejecting $H_0$.

3. Researchers commonly report the *p* value to be smaller or larger than alpha or mention that results are statistically significant rather than reporting exact *p* values. However, there is a difference between obtaining statistical significance with a *p* value that is slightly smaller than alpha as compared to when the *p* value is substantively smaller than alpha. The smaller the *p* value, the greater the evidence in favor of $H_a$. With commonly available online *p*-value calculators and spreadsheet functions, researchers can easily obtain exact *p* values and report them in research reports.

*Robert M. Capraro and Z. Ebrar Yetkiner*

***See also*** Effect Size, Measures of; Hypothesis; Sampling Distributions; Significance Level, Concept of; Significance Level, Interpretation and Construction

### Further Readings

Capraro, R. M. (2004). Statistical significance, effect size reporting, and confidence intervals: Best reporting strategies. *Journal for Research in Mathematics Education, 35,* 57–62.

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review, 48,* 378–399.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997–1003.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2003). *Applied statistics for the behavioral sciences* (5th ed.). Boston: Houghton Mifflin.

Kirk, R. E. (1982). *Experimental design: Procedures for the behavioral sciences* (2nd ed.). Belmont, CA: Brooks/Cole.

Mohr, L. B. (1990). *Understanding significance testing.* Newbury Park, CA: Sage.

Moore, D. S., & McCabe, G. P. (2006). *Introduction to the practice of statistics* (5th ed.). New York: W. H. Freeman.

Thompson, B. (2006). *Foundations of behavioral sciences: An insight-based approach*. New York: Guilford.

# ENCYCLOPEDIA OF RESEARCH DESIGN

# Editorial Board


## General Editor

Neil J. Salkind
*University of Kansas*

## Associate Editors

Bruce B. Frey
*University of Kansas*

Donald M. Dougherty
*University of Texas Health Science Center at San Antonio*

## Managing Editors

Kristin Rasmussen Teasdale
*University of Kansas*

Nathalie Hill-Kapturczak
*University of Texas Health Science Center at San Antonio*

# ENCYCLOPEDIA OF RESEARCH DESIGN

VOLUME 3

EDITED BY

NEIL J. SALKIND

*University of Kansas*

Los Angeles | London | New Delhi
Singapore | Washington DC

Copyright © 2010 by SAGE Publications, Inc.

All rights reserved. No part of this book may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission in writing from the publisher.

*For information:*

SAGE Publications, Inc.
2455 Teller Road
Thousand Oaks, California 91320
E-mail: order@sagepub.com

SAGE Publications Ltd..
1 Oliver's Yard
55 City Road
London EC1Y 1SP
United Kingdom

SAGE Publications India Pvt. Ltd.
B 1/I 1 Mohan Cooperative Industrial Area
Mathura Road, New Delhi 110 044
India

SAGE Publications Asia-Pacific Pte. Ltd.
33 Pekin Street #02-01
Far East Square
Singapore 048763

Printed in the United States of America.

*Library of Congress Cataloging-in-Publication Data*

Encyclopedia of research design/edited by Neil J. Salkind.
v. cm.
Includes bibliographical references and index.
ISBN 978-1-4129-6127-1 (cloth)
1. Social sciences—Statistical methods—Encyclopedias. 2. Social sciences—Research—Methodology—Encyclopedias.
I. Salkind, Neil J.

HA29.E525 2010
001.403—dc22 2010001779

This book is printed on acid-free paper.

10 11 12 13 14 10 9 8 7 6 5 4 3 2 1

| | |
|---|---|
| *Publisher:* | Rolf A. Janke |
| *Acquisitions Editor:* | Jim Brace-Thompson |
| *Editorial Assistant:* | Michele Thompson |
| *Developmental Editor:* | Carole Maurer |
| *Reference Systems Coordinators:* | Leticia M. Gutierrez, Laura Notton |
| *Production Editor:* | Kate Schroeder |
| *Copy Editors:* | Bonnie Freeman, Liann Lech, Sheree Van Vreede |
| *Typesetter:* | C&M Digitals (P) Ltd. |
| *Proofreaders:* | Kristin Bergstad, Kevin Gleason, Sally Jaskold, Sandy Zilka |
| *Indexer:* | Virgil Diodato |
| *Cover Designer:* | Glenn Vogel |
| *Marketing Manager:* | Amberlyn McKay |

# Contents

## Volume 3

List of Entries *vii*

Entries

| | | | |
|---|---|---|---|
| Q | *1149* | V | *1589* |
| R | *1183* | W | *1611* |
| S | *1295* | Y | *1645* |
| T | *1489* | Z | *1653* |
| U | *1583* | | |

Index *1675*

# List of Entries


Abstract
Accuracy in Parameter Estimation
Action Research
Adaptive Designs in Clinical Trials
Adjusted *F* Test. *See* Greenhouse–Geisser Correction
Alternative Hypotheses
American Educational Research Association
American Psychological Association Style
American Statistical Association
Analysis of Covariance (ANCOVA)
Analysis of Variance (ANOVA)
Animal Research
Applied Research
A Priori Monte Carlo Simulation
*Aptitudes and Instructional Methods*
Aptitude-Treatment Interaction
Assent
Association, Measures of
Autocorrelation

Bar Chart
Bartlett's Test
Barycentric Discriminant Analysis
Bayes's Theorem
Behavior Analysis Design
Behrens–Fisher *t′* Statistic
Bernoulli Distribution
Beta
Between-Subjects Design. *See* Single-Subject Design; Within-Subjects Design
Bias
Biased Estimator
Bivariate Regression
Block Design
Bonferroni Procedure
Bootstrapping
Box-and-Whisker Plot
b Parameter

Canonical Correlation Analysis
Case-Only Design
Case Study
Categorical Data Analysis
Categorical Variable
Causal-Comparative Design
Cause and Effect
Ceiling Effect
Central Limit Theorem
Central Tendency, Measures of
Change Scores
Chi-Square Test
Classical Test Theory
Clinical Significance
Clinical Trial
Cluster Sampling
Coefficient Alpha
"Coefficient Alpha and the Internal Structure of Tests"
Coefficient of Concordance
Coefficient of Variation
Coefficients of Correlation, Alienation, and Determination
Cohen's *d* Statistic
Cohen's *f* Statistic
Cohen's Kappa
Cohort Design
Collinearity
Column Graph
Completely Randomized Design
Computerized Adaptive Testing
Concomitant Variable
Concurrent Validity
Confidence Intervals
Confirmatory Factor Analysis
Confounding
Congruence
Construct Validity
Content Analysis
Content Validity
Contrast Analysis
Control Group
Control Variables
Convenience Sampling
"Convergent and Discriminant Validation by the Multitrait–Multimethod Matrix"
Copula Functions
Correction for Attenuation
Correlation
Correspondence Analysis
Correspondence Principle
Covariate
C Parameter. *See* Guessing Parameter
Criterion Problem
Criterion Validity
Criterion Variable
Critical Difference
Critical Theory
Critical Thinking
Critical Value
Cronbach's Alpha. *See* Coefficient Alpha

Crossover Design
Cross-Sectional Design
Cross-Validation
Cumulative Frequency Distribution

Databases
Data Cleaning
Data Mining
Data Snooping
Debriefing
Decision Rule
Declaration of Helsinki
Degrees of Freedom
Delphi Technique
Demographics
Dependent Variable
Descriptive Discriminant Analysis
Descriptive Statistics
Dichotomous Variable
Differential Item Functioning
Directional Hypothesis
Discourse Analysis
Discriminant Analysis
Discussion Section
Dissertation
Distribution
Disturbance Terms
*Doctrine of Chances, The*
Double-Blind Procedure
Dummy Coding
Duncan's Multiple Range Test
Dunnett's Test

Ecological Validity
Effect Coding
Effect Size, Measures of
Endogenous Variables
Error
Error Rates
Estimation
Eta-Squared
Ethics in the Research Process
Ethnography
Evidence-Based Decision Making
Exclusion Criteria
Exogenous Variables
Expected Value
Experience Sampling Method
Experimental Design
Experimenter Expectancy Effect
Exploratory Data Analysis
Exploratory Factor Analysis
Ex Post Facto Study
External Validity

Face Validity
Factorial Design
Factor Loadings
False Positive
Falsifiability
Field Study
File Drawer Problem
Fisher's Least Significant Difference Test
Fixed-Effects Models
Focus Group
Follow-Up
Frequency Distribution
Frequency Table
Friedman Test
*F* Test

Gain Scores, Analysis of
Game Theory
Gauss–Markov Theorem
Generalizability Theory
General Linear Model
Graphical Display of Data
Greenhouse–Geisser Correction
Grounded Theory
Group-Sequential Designs in Clinical Trials
Growth Curve
Guessing Parameter
Guttman Scaling

Hawthorne Effect
Heisenberg Effect
Hierarchical Linear Modeling
Histogram
Holm's Sequential Bonferroni Procedure
Homogeneity of Variance
Homoscedasticity
Honestly Significant Difference (HSD) Test
Hypothesis
Inclusion Criteria
Independent Variable
Inference: Deductive and Inductive
Influence Statistics
Influential Data Points
Informed Consent
Instrumentation
Interaction
Internal Consistency Reliability
Internal Validity
Internet-Based Research Method
Interrater Reliability
Interval Scale
Intervention
Interviewing
Intraclass Correlation
Item Analysis
Item Response Theory
Item-Test Correlation

Jackknife
John Henry Effect

Kolmogorov–Smirnov Test
KR-20
Krippendorff's Alpha
Kruskal–Wallis Test
Kurtosis

L'Abbé Plot
Laboratory Experiments
Last Observation Carried Forward
Latent Growth Modeling
Latent Variable
Latin Square Design
Law of Large Numbers
Least Squares, Methods of
Levels of Measurement
Likelihood Ratio Statistic
Likert Scaling
Line Graph
LISREL
Literature Review
*Logic of Scientific Discovery, The*
Logistic Regression
Loglinear Models
Longitudinal Design

Main Effects
Mann–Whitney *U* Test
Margin of Error
Markov Chains
Matching
Matrix Algebra
Mauchly Test
MBESS
McNemar's Test
Mean
Mean Comparisons
Median
Meta-Analysis
"Meta-Analysis of Psychotherapy Outcome Studies"
Methods Section
Method Variance
Missing Data, Imputation of
Mixed- and Random-Effects Models
Mixed Methods Design
Mixed Model Design
Mode
Models
Monte Carlo Simulation
Mortality
Multilevel Modeling
Multiple Comparison Tests
Multiple Regression
Multiple Treatment Interference
Multitrait–Multimethod Matrix
Multivalued Treatment Effects
Multivariate Analysis of Variance (MANOVA)
Multivariate Normal Distribution

Narrative Research
National Council on Measurement in Education
Natural Experiments
Naturalistic Inquiry
Naturalistic Observation
Nested Factor Design
Network Analysis
Newman–Keuls Test and Tukey Test
Nominal Scale
Nomograms
Nonclassical Experimenter Effects
Nondirectional Hypotheses
Nonexperimental Design
Nonparametric Statistics
*Nonparametric Statistics for the Behavioral Sciences*
Nonprobability Sampling
Nonsignificance
Normal Distribution
Normality Assumption
Normalizing Data
Nuisance Variable
Null Hypothesis
Nuremberg Code
NVivo

Observational Research
Observations
Occam's Razor
Odds
Odds Ratio
Ogive
Omega Squared
Omnibus Tests
One-Tailed Test
"On the Theory of Scales of Measurement"
Order Effects
Ordinal Scale
Orthogonal Comparisons
Outlier
Overfitting

Pairwise Comparisons
Panel Design
Paradigm
Parallel Forms Reliability
Parameters
Parametric Statistics
Partial Correlation
Partial Eta-Squared
Partially Randomized Preference Trial Design
Participants
Path Analysis
Pearson Product-Moment Correlation Coefficient
Percentile Rank
Pie Chart
Pilot Study
Placebo
Placebo Effect
Planning Research
Poisson Distribution
Polychoric Correlation Coefficient
Polynomials
Pooled Variance
Population
Positivism
Post Hoc Analysis
Post Hoc Comparisons
Power
Power Analysis
Pragmatic Study
Precision
Predictive Validity
Predictor Variable
Pre-Experimental Design
Pretest–Posttest Design
Pretest Sensitization
Primary Data Source
Principal Components Analysis
*Probabilistic Models for Some Intelligence and Attainment Tests*
Probability, Laws of
Probability Sampling
"Probable Error of a Mean, The"
Propensity Score Analysis
Proportional Sampling
Proposal
Prospective Study
Protocol
"Psychometric Experiments"
Psychometrics
Purpose Statement
*p* Value

Q Methodology
*Q*-Statistic
Qualitative Research
Quality Effects Model
Quantitative Research
Quasi-Experimental Design
Quetelet's Index
Quota Sampling

R
$R^2$
Radial Plot
Random Assignment
Random-Effects Models
Random Error
Randomization Tests
Randomized Block Design
Random Sampling
Random Selection
Random Variable
Range
Rating
Ratio Scale
Raw Scores
Reactive Arrangements
Recruitment
Regression Artifacts
Regression Coefficient
Regression Discontinuity
Regression to the Mean
Reliability
Repeated Measures Design
Replication
Research
Research Design Principles
Research Hypothesis
Research Question
Residual Plot
Residuals
Response Bias
Response Surface Design
Restriction of Range
Results Section
Retrospective Study
Robust
Root Mean Square Error
Rosenthal Effect
Rubrics

Sample
Sample Size
Sample Size Planning
Sampling
Sampling and Retention of Underrepresented Groups
Sampling Distributions
Sampling Error
SAS
Scatterplot
Scheffé Test
Scientific Method
Secondary Data Source
Selection
Semipartial Correlation Coefficient
Sensitivity
Sensitivity Analysis
Sequence Effects
Sequential Analysis
Sequential Design
"Sequential Tests of Statistical Hypotheses"
Serial Correlation
Shrinkage
Significance, Statistical
Significance Level, Concept of
Significance Level, Interpretation and Construction
Sign Test
Simple Main Effects
Simpson's Paradox
Single-Blind Study
Single-Subject Design
Social Desirability
Software, Free
Spearman–Brown Prophecy Formula
Spearman Rank Order Correlation
Specificity
Sphericity
Split-Half Reliability
Split-Plot Factorial Design
SPSS
Standard Deviation
Standard Error of Estimate
Standard Error of Measurement
Standard Error of the Mean
Standardization
Standardized Score
Statistic
Statistica
Statistical Control
*Statistical Power Analysis for the Behavioral Sciences*
Stepwise Regression
Stratified Sampling
Structural Equation Modeling
Student's *t* Test
Sums of Squares
Survey
Survival Analysis
SYSTAT
Systematic Error
Systematic Sampling

"Technique for the Measurement of Attitudes, A"
*Teoria Statistica Delle Classi e Calcolo Delle Probabilità*
Test
Test–Retest Reliability
Theory
Theory of Attitude Measurement
Think-Aloud Methods
Thought Experiments
Threats to Validity
Thurstone Scaling
Time-Lag Study
Time-Series Study
Time Studies
Treatment(s)
Trend Analysis
Triangulation
Trimmed Mean
Triple-Blind Study
True Experimental Design
True Positive
True Score
*t* Test, Independent Samples
*t* Test, One Sample
*t* Test, Paired Samples
Tukey's Honestly Significant Difference (HSD)
Two-Tailed Test
Type I Error
Type II Error
Type III Error

Unbiased Estimator
Unit of Analysis
U-Shaped Curve

"Validity"
Validity of Measurement
Validity of Research Conclusions

Variability, Measure of
Variable
Variance
Volunteer Bias

Wave
Weber–Fechner Law
Weibull Distribution
Weights
Welch's *t* Test
Wennberg Design
White Noise
Wilcoxon Rank Sum Test
WinPepi
Winsorize
Within-Subjects Design
Yates's Correction
Yates's Notation
Yoked Control Procedure

*z* Distribution
Zelen's Randomized Consent Design
*z* Score
*z* Test

# Q Methodology

Q methodology is a combination of conceptual framework, technique of data collection, and method of analysis that collectively provides the basis for the scientific study of subjectivity. This is distinguished from R methodology, which provides the basis for the study of what is objective in human behavior. Innovated in the mid-1930s by British physicist–psychologist William Stephenson, Q methodology focuses on opinions and perspectives that are gathered using the well-known Q-sort technique. These data are then submitted to factor analysis, pioneered by Stephenson's mentor Charles Spearman, which reveals the segmentation of subjectivity inherent in the substantive domain under consideration. Given the ubiquity of subjectivity, Q methodology applies to all areas of human endeavor—social attitudes, decision making, administration, the arts and humanities, cultural values, policy, economics, education, and so on, and including the natural and physical sciences insofar as subjective substructures undergird the theories and understandings used to explain objective facts. The focus on subjectivity in Q methodology requires departures from conventional applications of sampling, experimental design, and factor analysis. This entry provides information on the conceptual framework, data collection, and analysis of Q methodology, along with resources for additional information.

## Concourse of Subjective Communicability

Q methodology has its roots in *concourse,* a term coined by Stephenson that refers to the universe of subjective communicability surrounding any topic, of the kind found in ordinary conversation, back-fence gossip, commentary deposited on Internet blogs and exchanged in chat rooms, and extending to the high-level discourses of epistemic communities across all the sciences. Facts are invariably interlaced with opinions, and the division between the two turns on the principle of self-reference, its absence in R methodology and its centrality in Q. Hence, it is one thing to state that "free-flying wild birds can be a source of avian influenza," which is a matter of fact, and another to assert that "scientific surveillance of wild birds should be maintained worldwide," which is an opinion. The latter is self-referential to those who assert it, and it is apt to be accompanied by other opinions from the same concourse—for example, that "compartmentalization would reduce risk for introduction of AI viruses via trade," that "there is an immediate need to promote international research programs to develop better vaccines," and so forth.

The volume of opinion constitutes the universe of communicability that, in principle, is infinite in magnitude. Unlike sampling in surveys, where population boundaries can be specified and the number of cases is finite, the boundaries of communicability cannot be fixed and its content is limitless. Concourse is not restricted to statements of

opinion. The population of impressionist art also constitutes a concourse, as do the universes of novels, musical compositions, political cartoons, landscapes, and flavors of ice cream; in short, any collection of stimuli, linguistic or otherwise, for which individuals might express preferences. A defining feature of concourse is that its contents are generally untestable and incapable of falsification; however, they remain subject to measurement.

## Q-Sample Structuring and Experimental Design

Concourses are typically voluminous, as are person populations in survey research, and one of the steps in Q methodology involves reducing the concourse to a small sample of statements suitable for experimentation. As an illustration, consider the following statements, which were among more than 200 drawn from the media and interviews prior to the 2003 U.S. war with Iraq:

1. We should take this opportunity to free the Iraqi people from a dictator.
2. War will make more terror attacks likely.
3. I feel divided. On the one hand we have to take a stand so that the world knows we mean business. On the other hand I don't want America to start a big war.

As a counterpart to random sampling in survey research, the statement population in Q methodology is modeled theoretically (typically according to R. A. Fisher's experimental design principles), and statements are then selected in terms of the model. In this instance, examination of the concourse revealed statements compatible with the view of the U.S. administration (such as statement 1 above) and statements that were contrary to the regime's position (statement 2), but also statements that expressed ambivalence or reticence (statement 3), and virtually all of the statements in the concourse found easy placement in one of these three categories.

As a device for increasing diversity and breadth in the Q sample, all statements were reexamined and located in one of the three categories comprising Harold Lasswell's concept of *perspective:* For example, statement 1 expressed a *demand* (from a pro-regime standpoint), statement 2 an *expectation* (from an anti-regime standpoint), and statement 3 an *identification* (from a reticent point of view). All 200 statements were subsequently placed in one of the $(3)(3) = 9$ cells of the factorial design in Table 1, and $m = 5$ replicates from each cell were selected, for a Q-sample size of $N = (5)(9) = 45$ statements. Each of the statements was then typed on a separate card, resulting in a pack of 45 cards for subsequent Q sorting.

**Table 1** Q-Sample Structure (Iraq War)

| *Main Effects* | *Levels* | | | *n* |
|---|---|---|---|---|
| (*A*) Regime | (*a*) pro | (*b*) reticent | (*c*) anti | 3 |
| (*B*) Perspective | (*d*) demand | (*e*) identification | (*f*) expectation | 3 |

## Q Sorting

Q technique is the most widely known feature of Q methodology, and its popularity far exceeds awareness of the more abstract methodology that it was invented to serve. Q sorting consists of ranking the items in a Q sample according to some condition of instruction, typically from *mostly agree* to *mostly disagree*. Table 2 provides an example in terms of the Iraq War study; hence, this person agreed with statements 18, 22, and 45:

18. We have a professional army and the best military technology on the planet. The Iraqi army is no match.
22. The world would be safer without Saddam.
45. The United States has a duty to protect itself and the world.

and disagreed with statements 17, 30, and 40:

17. We're acting like a terrorist regime ourselves and only seem to be in this for self-interested capitalist motives.
30. This is a racist war, a campaign of genocide. We're just taking our 9/11 frustrations out on Iraq.
40. I must admit the idea of war scares me, and frankly I'd rather not hear about it.

This person obviously supports the impending conflict and at least manifestly embraces assertions emanating from the regime while denying with equal intensity propositions issued by counter elites.

The width of the range in Table 2 is relatively unimportant and has no statistical impact (nor does the shape of the distribution), but several principles concerning the Q sort are of considerable importance. First, as should be apparent, there is no correct way to perform the Q sort: It is a wholly subjective affair, which means that issues of validity are largely irrelevant. In a science of subjectivity, there can be no external criterion by which to validate "my point of view." Second, the distribution should ideally range from *mostly* to *mostly* (e.g., from *mostly agree* to *mostly disagree*), with *nonsalience* in the middle. This principle draws on 19th- and early 20th-century psychophysics, such as the work of Gustav Fechner and J. G. Beebe-Center, as well as observations dating from the earliest times down to Sigmund Freud's principle of pleasure-unpleasure and including the extremes of Likert scales. Third, meaning distends in opposite directions from zero and becomes most vivid at the extremes, as indicated in the tendency for Q sorters to feel most strongly about statements placed at the extremes and to be relatively nonengaged with respect to those located near zero. (Both of the above principles depend on a Q sample that is balanced in terms of the perspectives at issue; e.g., pro-regime vs. anti-regime.) Fourth, the forced distribution in Table 2 has the status of a *model,* in this case, a model of the Law of Error. Virtually all of the literature that has promoted an unforced (free) distribution has done so under the mistaken assumption that the forced distribution distorts how people would actually behave if unconstrained, rather than that it is an intentional constraint (although not without behavioral support) for inducing participants to reveal their preferences. Fifth, the statements in the Q sort are entangled; that is, they interact in a single Q-sort setting, and each is implicitly compared to, and achieves its score in relationship to, all others. (Entanglement is a signal feature of quantum mechanics, with which Q methodology demonstrates striking parallels.) This is in marked contrast to the situation in R methodology where a score obtained on a measured trait by any single individual is wholly independent of the scores obtained by other individuals for the same or any other trait. Sixth, the subjective character of both concourse and Q sorting necessitates (where feasible) a post-sorting interview so as to permit participants to elaborate upon their Q sorts; that is, to clarify the meanings that they attributed to the statements (which may be other than the meanings assigned in Table 1) and to justify the salience (from +4 to –4) that they assigned to each. This a posteriori determination of meaning is in marked contrast to R methodology, where scale meaning is achieved a priori in terms of validity and reliability tests that are carried out before responses are obtained.

**Table 2** Q-Sorting Distribution

| *Mostly Disagree* | | | | | | *Mostly Agree* | | |
|---|---|---|---|---|---|---|---|---|
| *–4* | *–3* | *–2* | *–1* | *0* | *+1* | *+2* | *+3* | *+4* |
| 17 | 29 | 3 | 9 | 4 | 2 | 1 | 8 | 18 |
| 30 | 32 | 13 | 11 | 7 | 6 | 5 | 12 | 22 |
| 40 | 35 | 21 | 15 | 10 | 14 | 24 | 26 | 45 |
| (3) | 37 | 43 | 23 | 16 | 20 | 31 | 28 | (3) |
| | 41 | 44 | 25 | 19 | 33 | 36 | 38 | |
| | (5) | (5) | 34 | 27 | 42 | (5) | (5) | |
| | | | (6) | 39 | (6) | | | |
| | | | | (7) | | | | |

Based on the 3 × 3 factorial design in Table 1, the Q sort in Table 2 could be analyzed in terms of the following variance components:

$$\Sigma d^2 = \Sigma A^2 + \Sigma B^2 + \Sigma AB^2 + \Sigma R^2$$

$$df(44) = (2) + (2) + (4) + (36),$$

where $\Sigma d^2$ is the Q-sort total sum of squares, $\Sigma A^2$ is the regime effect, $\Sigma B^2$ is the sum of squares due to perspective, $\Sigma AB^2$ is the regime × perspective interaction, and $\Sigma R^2$ is the replication variance (within-cell error). As noted at the beginning, however, the focus on subjectivity in Q methodology necessitates departures from conventional research assumptions, among them that items have an agreed-upon meaning as they do in scaling and as they would be required to have for the calculation

of means and variances necessary for the application of ANOVA. It is for this reason that the utility of variance designs is restricted to the composition of Q samples where they serve the important purpose of ensuring stimulus *representativeness,* a principle first advanced by Egon Brunswik. Beyond that, variance analysis yields to the operational principle and is superseded by correlation and factor analysis, which respond to the way in which participants actually operated with the statements rather than the way in which the investigator structured them.

It is to be noted that the sample of persons who perform the Q sort (P set) is also usually selected on the basis of experimental design procedures and is typically balanced or semibalanced for gender, age, party identification, or other salient variables arranged factorially. The P set is normally small (typically in the range of 30–50 participants), but as diverse as possible. The goal is for a representative set of participants to respond to a representative set of stimuli so as to maximize the likelihood that whatever subjective segmentations are in circulation will have an opportunity to reveal themselves.

## Correlation and Factor Analysis

Data analysis is illustrated in terms of an experiment on literary interpretation, an area of study rarely examined by other quantitative methods, but easily attended to using Q methodology due to the subjectivity involved. The focus in this instance is on the poem *Piano* by D. H. Lawrence, the reading of which produced a diversity of reactions that were subsequently reduced to a Q sample of size $N = 30$ critical comments including three levels of *reaction* (positive, negative, mixed) cross-classified with three kinds of *poetic concern* (emotion, technique, sense) in a $3 \times 3$ factorial arrangement. After reading the poem, eight graduate students of literature were instructed to provide their appraisals of the poem by Q sorting the statements from *agree* (+4) to *disagree* (–4), and Table 3 contains an abbreviation of the scores given the 30 statements by two of the students, *X* and *Y*.

**Table 3** Q-Sort Scores

| | $X$ | $Y$ | $d^2$ | $X^2$ | $Y^2$ |
|---|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 1 | 1 |
| 2 | –1 | 2 | 9 | 1 | 4 |
| 3 | 4 | 4 | 0 | 16 | 16 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 28 | 3 | –1 | 16 | 9 | 1 |
| 29 | 0 | 1 | 1 | 0 | 1 |
| 30 | –1 | 3 | 16 | 1 | 9 |
| $\Sigma$ | 0 | 0 | 82 | 182 | 182 |
| $\sigma$ | 2.46 | 2.46 | | | |

The formula for Pearson's correlation when means and standard deviations are the same (as they are when Q sorts adhere to the same forced distribution) is $r_{XY} = 1 - \dfrac{\Sigma d^2}{2N\sigma^2}$, where $N = 30$ statements, $\sigma^2$ is the variance of the Q-sort distribution, and $\Sigma d^2$ is the sum of squared differences in statement scores between the two Q sorts. Under these conditions, $2N\sigma^2$ is a constant and equivalent to the composite sums of squares of the two Q sorts ($182 + 182$ in this case); hence, the correlation between $X$ and $Y$ is $r_{XY} = 1 - \dfrac{82}{364} = .77$, which signifies a high degree of similarity between the two interpretations of the poem. The standard error of the correlation is given by $\sigma_{r=0} = \dfrac{1}{\sqrt{N}} = \dfrac{1}{\sqrt{30}} = .18$, and $2.58(.18) = .47$ indicates that correlations exceeding $\pm .47$ are significant ($p < .01$).

Given the $n = 8$ Q sorts, their intercorrelations result in an $8 \times 8$ correlation matrix, which is then factor analyzed using either of the two dedicated software packages, QMethod and PCQ (both of which can be accessed at www.qmethod.org), or by SPSS. Although principal components analysis (PCA) and varimax rotation are the most commonly used procedures in R methodology, investigators who use Q methodology are more inclined toward centroid factor analysis and increasingly toward theoretical rotation, although varimax is frequently used. Centroid analysis and theoretical rotation are both supported by the QMethod and PCQ programs, as are PCA and varimax.

The reasoning behind these analytic preferences again turns on the phenomenon of subjectivity and

the alteration in conventional analytic practices that this focus demands. The centroid solution is indeterminant (i.e., is one of an infinite number of solutions), and although varimax injects determinacy into the solution, it does so based on a statistical principle (simple structure) that is assumed to apply in any and all situations. Theoretical rotation, by way of contrast, takes context into consideration (in conformity with J. R. Kantor's principle of specificity) and as a result is increasingly relied upon as a way to find solutions that are *operant* (i.e., genuinely functional) as opposed to merely statistical. Implementation of theoretical rotation involves the graphical plotting of factor loadings in two-dimensional Cartesian space and manually rotating them so as to reveal suspected or hypothesized effects, and thereby relies on the abductory logic of Charles Peirce.

Table 4 shows the factor loadings for the eight participants, who offered three different interpretations of the poem, with Participant 1 advancing one interpretation (Factor A), Participants 2–5 advancing another (Factor B), and Participants 6–7 yet another (Factor C); Participant 8's response is mixed. The standard errors for correlations shown above apply to factor loadings, which are therefore significant in this case if in excess of $\pm 47$. The Q sorts comprising a factor are merged to provide a single Q sort, and this is effected by first weighting each response in terms of its factor loading: $w = \frac{f}{1 - f^2}$, where $f$ is the factor loading and $w$ is the weight. Consequently, in the case of Factor 2, Q sort 2 is weighted $w_2 = 1.26$ and Q sort 5 is weighted $w_5 = .82$, the latter therefore weighing only 65% of the former in the calculation of the factor scores. The factor scores are then calculated by multiplying each statement's Q-sort score by the weight and then summing each statement across the weighted Q sorts comprising the factor, with weighted statement sums then being converted into a factor array presented in the form of the original $+4$ to $-4$ metric.

**Table 4** Factor Loadings

| | *A* | *B* | *C* |
|---|---|---|---|
| 1 | **71** | 27 | –04 |
| 2 | –21 | **68** | –21 |
| 3 | –12 | **67** | 28 |
| 4 | 09 | **67** | 21 |
| 5 | –23 | **56** | –21 |
| 6 | –19 | –03 | **57** |
| 7 | –06 | 31 | **59** |
| 8 | **50** | **–52** | 28 |

*Note:* Decimals omitted; significant loadings are in boldface.

Examining Factor A, it is apparent that the persons comprising this factor have responded favorably to the poem, as shown in the following four statements (factor scores for A, B, and C, respectively):

| | | | |
|---|---|---|---|
| +4 | –2 | –2 | (*a*) The striking thing is that the poet knows quite well that this reversion to a childhood incident is sentimental, but he does not try to make capital out of the sentiment. |
| +4 | –2 | +4 | (*b*) One is made keenly aware of the strange relationship of past and present experience—one feels the emotion the poet experienced through his identity with and separation from his past self. |
| –4 | +3 | +2 | (*c*) The triviality of the sentiment is equaled only by the utter puerility of the versification. |
| –4 | –1 | +4 | (*d*) This poem is false. One worships the past in the present, for what it is, not for what it was. To ask for the renewal of the past is to ask for its destruction. |

The determination of significant differences among factor scores depends on factor reliability, which depends in turn on the reliability of the individual Q sorts. Adopting $\bar{r}_{1,2} = .80$ as a conservative estimate of average individual test-retest reliability, factor reliability is given by $r_{xx} = \frac{(.80)p}{1 + (p - 1).80}$, where $p$ is the number of persons defining the factor; hence, for Factor B, with $p = 4$ persons defining the factor, $r_{BB} = .94$. The standard error of factor scores is given by $\sigma_f = \sigma\sqrt{1 - r_{xx}}$, where $\sigma$ is the standard deviation of the Q-sort distribution (in this instance, from Table 3, $\sigma = 2.46$) and $\sigma_f$ is the standard error; hence, for Factor B, $\sigma_{fB} = .60$. Given that Factor A has only $p = 1$ defining Q sort, its factor reliability is $r_{AA} = .80$ and the standard error of its factor scores is $\sigma_{fA} = 1.10$. The

standard error of the difference in factor scores between Factors A and B is obtained by $\sigma_d = \sqrt{\sigma_{fA}^2 + \sigma_{fB}^2} = \sqrt{1.10^2 + .60^2} = 1.25$, which, when multiplied by 1.96 and rounded up to the nearest whole number, indicates that differences of 3 or more between scores for Factors A and B can be considered significant ($p < .05$). Based on this criterion, statements (*a*), (*c*), and (*d*) above can be seen to distinguish Factor A from the other two factors; statement (*b*) distinguishes Factor A from Factor B, but not from Factor C. Statistical information of this kind, which is routinely provided by the QMethod and PCQ software packages, along with postsorting interviews, comprise the main ingredients for the interpretation of factors.

## Factor Interpretation

Arguments and inferences in R methodology typically depend on the matrix of factor loadings, which reveals the connections among variables, but the factor loadings in Table 4 are of lesser interest in Q methodology inasmuch as the person sample is small and nonrepresentative by survey standards. Interest in Q methodology instead focuses on the array of factor scores, as with statements (*a*) to (*d*) above, because the factor scores reveal the subjectivity at issue, in this case concerning poetic interpretation. Interpretation is also more intricate in Q methodology because statements can, and frequently do, take on different meanings to different people, and to the same people in different settings, as opposed to scale items carrying meanings that have been established a priori.

In this regard, consider the Q-sort responses of 15 pediatric oncology patients concerning the fatigue that they experienced during treatment. The responses resulted in three factors, the first of which (Factor X) assigned positive scores to the following distinguishing statements (scores in parentheses for Factors X, Y, and Z, respectively):

> I had energy (+4, 0, –3). I felt great (+4, +2, –4). I had enough energy to play as hard as I wanted (+3, –3, –3). I was so tired that I needed to rest during the day (+2, 0, –1). I had enough energy to play sports (+2, –2, –4).

The statement about needing rest provided a clue indicating that these patients had incorporated a rest period into their regimen that enabled them to regenerate and lead otherwise normal lives. Factor Y, on the other hand, told a different story (scores for Factors X, Y, and Z, respectively):

> Being tired made it hard to keep up with my friends (–1, +4, +1). Being tired made it hard to concentrate (+1, +4, +1). Being tired made me sad (–3, +3, –3). My tiredness was hard on my family (–2, +2, –1). I was too tired to watch television (–1, –4, 0). I felt sleepy (+3, –2, +3).

The relative emphasis on concentration and sadness led to a conclusion that fatigue was affecting these patients in cognitive, emotional, and social ways that might require counseling to reactivate them and get them out from in front of the television. Note, however, that their fatigue did not spill over into sleepiness. Finally, Factor Z:

> I feel tired (+2, –1, +4). I felt weak (+1, +1, +4). Being tired made me mad (–3, +1, +3). I had trouble finishing things because I was too tired (0, –1, +2).

These patients felt weak and not just tired, and unlike Factor Y, their emotional reaction was one of anger rather than sadness, hence providing them with a motive force for improvement that Factor Y lacked.

The above are bare caricatures of the three factors, which contained interesting nuances that could be drawn out only through careful and prolonged examination of the factor arrays. The results led to suggestions for crafting a questionnaire, based on the Q factors, that could serve as a screening device aimed at distinguishing patients who were experiencing one of these three reactions to fatigue. Factor X patients already have control over their lives and require little in the way of assistance, whereas the Factor Y and Factor Z patients require different therapeutic interventions. The interpretive phase of a Q study has much in common with hermeneutics as well as narrative and discourse analysis, and it is at this point that Q methodology resembles qualitative methods, despite its roots in variance design and factor theory. As a practical matter, the factors that it produces represent the vital signs of human

subjectivity and serve as decision structures that point in promising directions of implementation.

### Resources

Additional information about Q methodology can be found at www.qmethod.org, including information on an electronic discussion group and the International Society for the Scientific Study of Subjectivity. Three journals specialize in Q-related studies: *Operant Subjectivity, Journal of Human Subjectivity,* and the Korean language *Q-Methodology and Theory.* Books on Q methodology are also now available in Korean, Persian, Romanian, and Thai.

*Steven R. Brown and James M. M. Good*

***See also*** Correlation; Experimental Design; Factor Loadings; Qualitative Research; Variance

#### Further Readings

Brown, S. R. (1980). *Political subjectivity: Applications of Q methodology in political science.* New Haven, CT: Yale University Press.

Brown, S. R., Durning, D. W., & Selden, S. C. (2008). Q methodology. In K. Yang & G. J. Miller (Eds.), *Handbook of research methods in public administration* (2nd ed., pp. 721–763). New York: Taylor & Francis.

Brown, S. R., & Robyn, R. (2004). Reserving a key place for reality: Philosophical foundations of theoretical rotation. *Operant Subjectivity, 27,* 104–124.

McKeown, B. F., & Thomas, D. B. (1988). *Q methodology.* Newbury Park, CA: Sage.

Stenner, P., Watts, S., & Worrell, M. (2007). Q methodology. In C. Willig & W. Stainton Rogers (Eds.), *The Sage handbook of qualitative research in psychology* (pp. 215–239). London: Sage.

Stephenson, W. (1953). *The study of behavior: Q-technique and its methodology.* Chicago: University of Chicago Press.

#### Websites

Q Methodology: http://www.qmethod.org

## Q-Statistic

*Q*-statistic is a nonparametric inferential test that enables a researcher to assess the significance of the differences among two or more matched samples on a dichotomous outcome. It can be applicable in a situation in which a categorical variable is defined as success and failure. The data are distributed in a two-way table; each column, *j,* represents a sample and each row, *i,* a repeated measure or a matched group. Thus, the *Q*-test is

$$Q = \frac{k(k-1)\sum_{j=1}^{k}\left(T_j - \overline{T}\right)^2}{k\sum_{i=1}^{n}u_i - \sum_{i=1}^{n}u_i^2}, \qquad (1)$$

where $T_j$ is the total number of successes in the *j*th sample (column), $u_i$ is the total number of successes in the *i*th row, and $k$ is the total number of samples. A most convenient equation to compute the *Q*-statistic is

$$Q = \frac{(k-1)\left[k\sum_{j=1}^{k}T_j^2 - \left(\sum_{j=1}^{k}T_j\right)^2\right]}{k\sum_{i=1}^{n}u_i - \sum_{i=1}^{n}u_i^2}. \qquad (2)$$

Cochran developed this test under a permutation model because the population of possible results in the *i*th row consists of the $\binom{k}{u_i}$ different combinations of the $u_i$ successes among the $k$ columns, and their results pertaining to the null case are asymptotic (Blomqvist, 1951). If the true probability of success is the same in all samples, the limiting distribution of *Q,* when the number of rows is large, is the $\chi^2$ distribution with $k - 1$ degrees of freedom because the joint distribution of the column totals $T_j$ may be expected to tend to a multivariate normal distribution with a common variance-covariance matrix.

### Approximations to Q-Statistic

Usually, the $\chi^2$ test is used to make inferences about ratios or percentages, but when the samples are correlated, the use of this test violates the assumption of the independence among the samples compared.

The *Q*-test is an inferential statistic developed by the statistician William Gemmell Cochran (1909–1980) that arises as an extension of the McNemar test.

The McNemar test examines the significance of the differences between ratios or percentages of two correlated samples. The $Q$-statistic allows for the evaluation of the null hypothesis of equality between ratios or percentages for more than two matched samples under a permutation model, so it can be simplified to McNemar's test when there are only two samples.

The $Q$-statistic is also equivalent to the sign test when the samples are small and there are no significant outliers. The sign test was developed for testing the median difference of independent paired observations.

The distribution of the $Q$-test for small samples has been approximate to the $\chi^2$ and $F$ test using a correction for continuity. However, rows containing only 1s or 0s can yield quite different results using the $F$ approximation without affecting the value of the $Q$-statistic. Although the $F$ test, corrected for continuity, can get a better approximation than the corrected or not $\chi^2$ in the same cases, the latter has been taken as the common approximation because it is easier to calculate. Nevertheless, as those computations are run by computers today, the possible application of the corrected $F$ approximation should be considered.

The accuracy of the $Q$-statistic in small samples depends on the number of conditions as well as the sample size; the $\chi^2$ approximation seems good enough with a total of 24 scores or more, deleting those rows with only 1s or 0s. After deleting those rows without variation, if the total product, columns by rows, is less than 24, the exact distribution should be constructed. The distribution of $Q$-test can be obtained by the method that Patil proposed in 1975 or by Tate and Brown's 1964 $r \times c$ tables, where the probabilities vary in dimensions from 3 columns and 12 rows to 6 columns and 4 rows.

Finally, the $Q$-statistic has been also tested under a model for dichotomous responses obtained in experiments within which randomization is not necessarily incorporated, and it remains appropriate for testing the null hypothesis that all measured conditions yield the same percentage of successes.

## Example

A hypothetical example is presented here to make clearer the context and how the $Q$-statistic is used. Imagine we want to examine the effect of an intervention to give up smoking. A sample of $n = 8$ subjects was evaluated immediately after the intervention and in three more time periods, $k = 4$, to assess the maintenance effect of the intervention. At each assessment ($T_j$), the participant's response was dichotomized as smoking or not, and giving up smoking ($u_{ij}$) was considered the success. Their distribution is presented in Table 1.

**Table 1** Effect of the Intervention for Eight Individuals to Give Up Smoking at Four Evaluations Across Time

| | *Time Period* | | | | |
|---|---|---|---|---|---|
| *Individual* | $T_1$ | $T_2$ | $T_3$ | $T_4$ | *Total*, $u_i$ |
| 1 | 1 | 0 | 0 | 0 | 1 |
| 2 | 1 | 1 | 1 | 0 | 3 |
| 3 | 1 | 1 | 1 | 1 | 4 |
| 4 | 1 | 1 | 0 | 0 | 2 |
| 5 | 1 | 0 | 0 | 0 | 1 |
| 6 | 1 | 0 | 0 | 0 | 1 |
| 7 | 1 | 1 | 1 | 0 | 3 |
| 8 | 1 | 1 | 1 | 0 | 3 |
| Total, $T_j$ | 8 | 5 | 4 | 1 | $\sum_{1}^{32} u_{ij} = 18$ |

In this example, the four times that the participants were evaluated are like four samples with the same group of subjects (repeated measures), so either this kind of design or one in which there are different groups of matched individuals are considered correlated samples. The $Q$-statistic is used in those cases to test the hypothesis that the percentage of successes does not change across samples or, in our example, across time

$$H_0 : P(T_1) = P(T_2) = \cdots = P(T_j),$$

versus the alternative hypothesis that at least two of the samples present statistically significant differences.

Following our example, and considering the total data in Table 1 with $n = 8$, $k = 4$, and row and column total as shown, the differences across time can be tested:

$$\sum_{j=1}^{k=4} T_j^2 = 8^2 + 5^2 + 4^2 + 1^2 = 106,$$

and

$$\sum_{i=1}^{n=8} u_i^2 = 1^2 + 3^2 + 4^2 + \cdots + 3^2 + 3^2 = 50.$$

Substituting in Equation 2, we obtain

$$Q = \frac{3\left[4(106) - 18^2\right]}{4(18) - 50} = 13.64.$$

The $Q$ value obtained with $k - 1 = 7$ degrees of freedom has a probability in its approximation of a $\chi^2$ distribution (because the total of scores is higher than 24) of $p = .003$. Therefore, it is possible to conclude that the effectiveness of the intervention differs significantly across time. After this analysis, some post hoc multiple comparisons can be carried out to analyze the linear, quadratic, and/or cubic trends across time or across the matched samples.

Another design for which we will apply this technique is when the groups are matched and it is necessary to know if it is possible to assume homogeneous percentages of the dichotomy response among the matched groups. Matched groups is an experimental method for controlling different individual characteristics (e.g., age, sex, race, socioeconomic status) because of their possible influence on the results. Thus, the subjects in one group are matched on a one-to-one basis with subjects in other groups concerning those different individual characteristics that have been previously tested.

Several software packages (e.g., IBM® SPSS® Statistics, formerly called PASW® Statistics, and Stata) contain this statistic, and usually the input is the columns for each sample with the outcome, coded with 0s and 1s, and the output is the $Q$-statistic, number of subjects, number of repeated measures or matched groups, degrees of freedom, and upper-tail $p$ value.

## The Variability of a Combination of Estimates From Different Experiments

It is important to note that a statistic also called $Q$-statistic was developed by Cochran in 1954 in order to have a test of significance of the variation in response among effect sizes from different experiments. Thus, it is usually used to assess whether there is true heterogeneity in research synthesis methodology.

The $Q$-test is computed by summing the squared deviations of each study's effect estimate from the overall effect estimate, weighting the contribution of each study by its inverse variance as

$$Q = \sum w_i \left(T_i - \overline{T}\right)^2, \tag{3}$$

where $w_i$ is the weighting factor, inverse of the estimated variance of the effect size, $\hat{\sigma}_i^2$, for the $i$th study assuming a fixed-effects model ($w_i = 1/\hat{\sigma}_i^2$), and $\overline{T}$ is

$$\overline{T} = \frac{\sum_i w_i T_i}{\sum_i w_i}. \tag{4}$$

If we assume that the conditional within-study variances, $\sigma_i^2$, are known, under the null hypothesis of homogeneity ($H_0 : \delta_1 = \delta_2 = \ldots = \delta_k$; or also $H_0 : \tau^2 = 0$), the $Q$-statistic has a $\chi^2$ distribution with $k - 1$ degrees of freedom, $k$ being in this case the number of integrated studies. Thus, $Q$ values higher than the critical point for a given significance level ($\alpha$) enable us to reject the null hypothesis and conclude that there is statistically significant between-study variation. If the homogeneity assumption is not met, the meta-analyst has to look for moderator variables to explain the heterogeneity; sometimes, assuming a random-effects model, which includes the within- and between-study variability, allows one to accept the null hypothesis.

A weakness of the $Q$-statistic is that it has poor power to detect true heterogeneity among studies when the meta-analysis includes a small number of studies and excessive power to detect negligible variability with a high number of studies.

## Example for Q-Statistic in Meta-Analysis

Some data sets from a meta-analysis about the efficacy of tailored intervention through computers for health behavior are introduced here. The treatment and control groups were compared using the standardized mean difference, $d$, as an effect size

**Table 2** Effect Size, 95% CI, Variances, and Weights of Tailored Interventions for Health Behavior

| *Study* | *d (95% CI)* | $\hat{\sigma}_i^2$ | $w_i$ |
|---|---|---|---|
| 1 | 0.25 (0.13, 0.37) | 0.0147 | 68.1446 |
| 2 | 0.12 (0.00, 0.23) | 0.0136 | 73.7149 |
| 3 | 0.22 (0.10, 0.34) | 0.0152 | 65.8396 |
| 4 | 0.47 (0.40, 0.54) | 0.0048 | 206.9776 |
| 5 | 0.56 (0.48, 0.63) | 0.0051 | 196.6882 |
| 6 | 0.13 (0.06, 0.20) | 0.0047 | 213.3799 |
| 7 | 0.05 (–0.02, 0.11) | 0.0046 | 219.3887 |
| 8 | 0.12 (0.05, 0.18) | 0.0045 | 222.8813 |

metric ($T$). The effect size values for eight studies are presented in Table 2.

Substituting in Equations 4 and 3 the values of our example,

$$\overline{T} = \frac{(68.1446 \times 0.25) + \cdots + (222.8813 \times 0.12)}{68.1446 + \cdots + 222.8813} = 0.24$$

$$Q = \left(68.1446 \times (0.25 - 0.24)^2\right) + \cdots + \left(222.8813 \times (0.12 - 0.24)^2\right) = 45.66,$$

it is possible to conclude that the distribution of the effect sizes is not homogeneous because the probability of the $Q$ value with 7 degrees of freedom is lower than .05.

*Tania B. Huedo-Medina*

***See also*** Effect Size, Measures of; *F* Test; McNemar's Test; Meta-Analysis; Sign Test

### Further Readings


Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics, 10,* 101–129.

Hardy, R. J., & Thompson, S. G. (1998). Detecting and describing heterogeneity in meta-analysis. *Statistics in Medicine, 17,* 841–856.

Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: $Q$ statistic or $I^2$ index? *Psychological Methods, 11,* 193–206.

McNemar, W. (1949). *Psychological statistics*. New York: Wiley.

Patil, K. D. (1975). Cochran's $Q$ test: Exact distribution. *Journal of the American Statistical Association, 70,* 186–189.

Tate, M. W., & Brown, S. M. (1964). *Tables for comparing related-sample percentages and for the median test* [Available on microfiche]. Graduate School of Education, University of Pennsylvania.


# Qualitative Research

Qualitative research, also known as qualitative inquiry, is an umbrella term used to cover a wide variety of research methods and methodologies that provide holistic, in-depth accounts and attempt to reflect the complicated, contextual, interactive, and interpretive nature of our social world. For example, grounded theory, ethnography, phenomenology, ethnomethodology, narratology, photovoice, and participatory action research (PAR) may all be included under the qualitative label, although each of these individual methods is based on its own set of assumptions and procedures. What unifies these various approaches to inquiry is their primary reliance on non-numeric forms of data (also known as empirical evidence) and their rejection of some of the underlying philosophical principles that guide methods employed in the physical and natural sciences and frequently in the social sciences. This entry focuses on the philosophical frameworks positioning qualitative research and on qualitative research designs.

## Positioning Qualitative Inquiry

At the heart of the distinction between these various scientific methods are differing ontological, epistemological, and theoretical worldviews. Ontology refers to the nature of reality. Ontological questions interrogate fundamental ideas about what is real. Epistemology refers to a theory of knowledge. Epistemological discussions interrogate how we know the world, who can know, and what can be known. Theoretical perspectives are the philosophical stances that provide the logic and the criteria that organize methodology (the overall research strategy) and methods (the specific tools or techniques used in collecting and interpreting evidence). In short, basic philosophical differences

in these worldviews have a direct impact on the research design. Coherent research designs demonstrate consistent and integrated ontological, epistemological, theoretical, and methodological positions.

Most qualitative research starts from a constructivist epistemological position and from one of a variety of theoretical perspectives, such as interpretivist, feminist, or critical inquiry. Constructivists believe in the socially constructed nature of reality. This idea that reality is generated through social interaction and iterative processes has dramatic implications for addressing the basic epistemological questions raised above and thus for the methodologies and methods employed. Constructivists reject the basic premise that an objective researcher discovers truths from preexisting data. Instead, they believe in what Norman Denzin and Yvonna Lincoln have called an "intimate relationship" between the researchers and the phenomenon under investigation. Marianne Phillips and Louise Jorgensen argue, based on the work of Vivien Burr and Kenneth Gergen, that constructivist approaches share four basic premises: a critical approach to taken-for-granted knowledge that is often overlooked or ignored, an interest in historical and cultural specificity, a link between knowledge development and processes, and a link between knowledge development and social action.

Within the constructivist epistemological tradition, there are many different theoretical schools of thought. Interpretivism has spawned a family of related traditions such as symbolic interactionism, phenomenology, and hermeneutics. For example, in his classic work on symbolic interactionism in 1969, Herbert Blumer summarized the three basic premises underlying his work: Humans act toward things on the basis of meaning that those things have for them; meaning is derived from, or arises out of, social interaction with others; and meanings attach and are modified through an interpretative process during interactions. Note that at the heart of symbolic interactionism is a belief that reality is not stable and pre-existing but rather is socially constructed and given meaning only through ongoing interactions. Critical theorists believe that social action and social reform are an integral part of any research endeavor. Today's critical inquiry was heavily influenced by the 1972 work of Paulo Freire, *Pedogogy of the Oppressed,* which fueled important discussions about power relationships, oppression, exploitation, empowerment, democracy, social justice, and action-based research. Feminists in the 1970s and beyond have furthered these theoretical discussions, raising basic questions such as those posed by Sandra Harding in the title of her seminal book published in 1991, *Whose Science? Whose Knowledge?*

In general, qualitative researchers are interested in studying social processes, how people make sense and create meaning, and what their lived experiences are like. They are interested in understanding how knowledge is historically, politically, and culturally situated. They concern themselves with notions of power, privilege, positionality, and social justice. They are likely to acknowledge research as a value-laden activity and embrace the idea that the researcher himself or herself is an instrument in the process. Often, they try to recreate experiences, understanding, and meanings from the point of view of those being studied rather than positioning themselves as the final interpretative authority or expert.

Qualitative research is sometimes contrasted with quantitative research. It is worth considering the distinctions. Most quantitative endeavors start from an objectivist epistemological perspective and a positivist theoretical position. Because quantitative analysis and statistical techniques are consistent with positivist assumptions about both the nature of social reality (an ontological issue) and the relationship between the "knower" and the "known" (epistemological issues), they tend to dictate how studies are designed and evaluated. Positivists are apt to believe in a stable reality and attach a premium to the objectivity, neutrality, and expertise of the scientist studying it. The research process is considered value-neutral. Variables or standardized instruments are used to measure phenomena and test relationships. Research designs must be carefully constructed in order to ensure that the mathematical assumptions that underlie statistical tests are met. The goal is often to generalize findings to larger populations.

Constructivist critiques of the objectivist perspective include concerns that these positivist methodologies tended to ignore or minimize contexts

(such as political, historical, or cultural factors); they ignored the meanings people attached to their actions or experiences; they were divorced from real-world settings; they focused on generalizable results that were of little use to individual decision makers; they tended to favor testing pre-existing hypotheses and theories rather than generating new ones; they favored deductive logic over inductive approaches; and they were reductionistic rather than holistic and thus did not adequately account for the complexity of social phenomena. An oft-repeated characterization holds that quantitative research is a mile wide and an inch deep, whereas qualitative research is a mile deep and an inch wide.

In short, it is critical to understand that qualitative inquiry starts with a very different relationship to social reality than that of most quantitative researchers. Constructivists challenge positivist views about the nature of reality, how to study it, who can know, what knowledge is, and how it is produced. This influences every aspect of the research design. So qualitative research will look markedly different—at every stage of the design and implementation process—than quantitative research.

## Qualitative Research Designs

Because a variety of different methods fall within the scope of qualitative research, it is difficult, even dangerous, to make generalizations about designs. Readers should consult methodologists who specialize in a particular form of inquiry to learn about the rules of that particular method. That said, what follows is a summary of features that tend to be associated with qualitative inquiry.

### *Overall Design*

Qualitative designs are emergent and flexible, standing in stark contrast to quantitative research designs in which a hallmark feature is their fixed and predetermined nature. Qualitative work is sometimes characterized as "messy," not because the researcher is sloppy but because the process is not strictly controlled and must be adapted to the realities of the environment. Designs often rest on inductive approaches to knowledge development. They may be value-driven, acknowledge the politics of the research endeavor, be attentive to power differentials in the community, and promote social justice or social change.

### *Role of the Researcher*

In qualitative inquiry, the researcher is often called an "instrument" or "tool" in the process. This characterization acknowledges that all interpretations and observations are filtered through the researcher, who brings his or her own values and identity to the process. Qualitative researchers talk about situating themselves relative to the research project and participants. For example, they may be insiders (share characteristics with those studied) or outsiders or some combination of both. Furthermore, the researcher's role can fall anywhere on a continuum of action from unobtrusive observer to full participant. Many designs call for the researcher to be a participant-observer. Action-based approaches to inquiry tend to characterize the role of the researcher as collaborator.

In short, qualitative research designs are more apt to require the researcher to make his or her position, role, and influence transparent. This is unlike approaches that try to minimize or neutralize the researcher's presence. Qualitative researchers are expected to critically examine and disclose their position. They do this by being "self-reflexive." Michelle Fine has written about "working the hyphen," a metaphor that is useful in acknowledging actors on both sides of the researcher–participant equation.

### *Research Question*

The framing of a qualitative research question will depend on the method. However, such questions tend to be both broad and flexible. They are not variable-driven, so they do not seek to link concepts and posit relationships. A common feature is that the research question evolves during the process. For example, Elliot Liebow, an ethnographer, started his study on homeless women by asking the research question: How do homeless women see and experience homelessness? His initial position was that he knew nothing about what their lives were like. However, in his final report, he ended up asking and answering the question

"How do homeless women remain human in the face of inhuman conditions?" This occurred because during the process of his data collection and analysis, he learned enough about homeless women's experiences to answer a more sophisticated, complicated, and interesting research question about their lived experiences.

### *Study Site*

Qualitative studies most often take place in natural environments. Researchers want to observe how people live and act in the real world, not when they are removed from these settings. For this reason, they are likely to study people and things "in the field," which can refer to any natural environment such as social agencies, prisons, schools, rural communities, urban neighborhoods, hospitals, police stations, or even cyberspace.

### *Study Sample*

Qualitative researchers use very different sampling strategies from quantitative researchers. Large, random, representative samples are rarely, if ever, the goal. Qualitative researchers are more apt to use some form of purposive sampling. They might seek out people, cases, events, or communities because they are extreme, critical, typical, or atypical. In some qualitative methods, sampling may be based on "maximum variation." So, the idea is to have as diverse a pool on the characteristics under investigation as possible. These studies are likely to be exploring the breadth and depth of difference rather than similarities. Sampling may be theoretically driven. For example, grounded theory is a method that seeks to generate mid-level theory as a research product. For this reason, it uses theoretical sampling strategies, picking new cases on the basis of the emerging theory.

### *"Data" or Empirical Evidence Collection*

In general, the empirical evidence used in qualitative inquiry is non-numeric and is collected in one or more of three basic forms: through *interviews and/or conversations* (either one-on-one or in groups), through *observations* (either unobtrusive or as a participant), and/or through *documents and artifacts* (either pre-existing or generated as part of the research process).

*Interviews* can be unstructured, semi-structured, or structured (although this is rarely favored). Questions are usually open-ended and seek to give the participant an opportunity to answer fully. Their purpose is often to explore meaning, understanding, and interpretations rather than to treat the interviewee as a vessel for retrieving facts. Interviews can be conducted with multiple individuals, such as in focus groups. They are often audiotaped or videotaped. These tapes can then be transcribed (using any number of different formats and transcription techniques) for easier analysis.

*Observations* occur in natural settings. The researchers must spend considerable time recording observations (through memos, jottings, and field notes) to preserve them for future use. It is critical that these notes be written during field observation or as soon thereafter as possible while they are fresh in mind because they serve as the empirical evidence for use during analysis.

*Documents or artifacts* used in qualitative work can be pre-existing, such as newspaper articles, client case files, legislative documents, or policy manuals and are collected by the researcher. Alternatively, documents or artifacts may be generated as part of the research process itself, such as asking subjects to keep journals, take photographs, or draw pictures. For example, in a neighborhood safety study, children might be asked to draw a map of their route between home and school and to mark the scary places. In an action-based method known as photovoice, participants take community-based photographs that are then used as a springboard for advocacy efforts.

### *Empirical Evidence Analysis*

Obviously, the form of empirical evidence analysis must be appropriate for the methods employed. Some common methods of qualitative analysis include thematic analysis, constant comparative methods, discourse analysis, content analysis, and narrative analysis. For example, some researchers conduct thematic analysis where they closely examine their texts (interviews, field notes, documents), apply codes, and develop themes. Researchers engaged in narrative analysis look at the way life stories are structured, such as the

divorce narratives studied by Catherine Reissman. Discourse analysis seeks to reveal patterns in speech. Similarly, conversational analysis looks closely at the structure, cadence, and patterns in ordinary conversations such as Douglas Maynard's study on how "good news" and "bad news" are delivered in clinical medical settings. Ethnographers review their field notes, interview transcripts, and documents and may take an interpretative approach to analysis. Action-based research methods often call for the study participants to play an active role in co-constructing information, so the analysis and interpretation do not reside solely in the hands of the "expert" researcher. Many qualitative researchers employ a procedure known as "member checking," where they take their preliminary interpretations back to participants to check whether those interpretations ring true.

Some qualitative researchers use computer software programs to aid them in their work; others do not. There are a number of different software packages available, such as HyperResearch, NVivo, Atlas.ti, QDA Miner, and AnSWR, among others. They can be useful in helping the researcher manage, organize, and retrieve large amounts of text, videos, images, and other forms of qualitative evidence. However, unlike statistical software, it is important to recognize that qualitative programs are not designed to "analyze" data itself. Virtually every aspect of qualitative analysis process relies heavily on the interpretative and analytic procedures carried out by the researcher.

### *Findings and Writing Reports*

Qualitative research reports often read very differently from traditional scientific reports. For example, it is common for qualitative researchers to use the first person in their writing rather than the third person often found in quantitative reports. These voice preferences are traceable to the epistemological roots discussed above. First-person narrations may appear odd to those trained to read and think about the researcher as a detached, objective, neutral observer, but they are consistent with epistemological and theoretical worldviews that posit that the researcher plays an active role in the process. Furthermore, because qualitative researchers do not attempt to measure phenomena, they rarely report quantity, amounts, intensity, or frequency and are more likely to present their findings as complicated and detailed narratives. In 1973, Clifford Geertz first introduced the term *thick description* to describe the rich and contextual writing that is often the product of ethnography or other cultural studies. However, this can be controversial. Norman Denzin and Yvonna Lincoln have noted the "crisis in representation" that first occurred among anthropologists, who worried about the politics and ethics of their textual representation of "others." Qualitative researchers often speak of "giving voice" to their study participants and attempt to stay close to the words and meanings of their informants.

### *Quality and Rigor in Qualitative Research*

It is important to evaluate the quality of qualitative studies relative to standards that are applicable to the specific research method used. Sometimes, qualitative researchers refer to negotiated validity, trustworthiness, transferability, transparency, and credibility as quality indicators. For example, Joseph Maxwell has argued for a broad understanding of validity with multiple dimensions such as descriptive validity, interpretive validity, theoretical validity, internal generalizability, and evaluative validity. Yvonna Lincoln and Egon Guba argued for judging case studies based on resonance, rhetoric, empowerment, and applicability.

*Karen M. Staller*

*See also* Action Research; Critical Theory; Discourse Analysis; Ethnography; Focus Group; Grounded Theory; Interviewing; Narrative Research; Naturalistic Observation; NVivo; Observational Research; Observations

**Further Readings**

Agar, M. H. (1980). *The professional stranger: An informal introduction to ethnography.* New York: Academic Press.

Aunger, R. (2004). *Reflexive ethnographic science.* Walnut Creek, CA: Alta Mira.

Brown, L., & Strega, S. (2005). *Research as resistance: Critical, indigenous, & anti-oppressive approaches.* Toronto, Ontario: Canadian Scholars' Press.

Clandinin, D. J., & Connelly, F. M. (2000). *Narrative inquiry: Experience and story in qualitative research*. San Francisco: Jossey-Bass.
Crotty, M. (1998). *The foundations of social research: Meaning and perspectives in the research process*. Thousand Oaks, CA: Sage.
Denzin, N. K., & Lincoln, Y. S. (2005). *Handbook of qualitative research* (3rd ed.). Thousand Oaks, CA: Sage.
Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory*. Chicago: Aldine.
Hesse-Biber, S., & Leavy, P. (2004). *Approaches to qualitative research: A reader on theory and practice*. New York: Oxford University Press.
Lincoln, Y., & Guba, E. G. (1985). *Naturalistic inquiry*. Beverly Hills, CA: Sage.
Polkinghorne, D. E. (1988). *Narrative knowing and the human sciences*. Albany: State University of New York Press.
Reissman, C. K. (1993). *Narrative analysis*. Newbury Park, CA: Sage.

# Quality Effects Model

The quality effects model is a method of statistical synthesis of data from a set of comparable studies of a problem, and it yields a quantitative summary of the pooled results. It is used in a process called meta-analysis where there is an aggregating of the data and results of a set of studies. The process is widely used in the biomedical sciences, especially in epidemiology and in clinical trials.

## Weighting by Inverse Variance

Because the results from different studies investigating different independent variables are measured on different scales, the dependent variable in a meta-analysis is some standardized measure of effect size. One of the commonly used effect measures in clinical trials is a relative risk (RR), when the outcome of the experiments is dichotomous (success vs. failure). The standard approach frequently used in meta-analysis in clinical research is termed the *inverse variance method* or *fixed-effects model* based on Woolf and first proposed by Gene Glass in 1976. The average effect size across all studies is computed, whereby the weights are equal to the inverse variance of each study's effect estimator. Larger studies and studies with less random variation are given greater weight than smaller studies. In the case of studies reporting a RR, the logRR has a standard error (*se*) given by

$$se_i = \sqrt{\frac{(1 - P_{iT})}{P_{iT} n_{iT}} + \frac{(1 - P_{iC})}{P_{iC} n_{iC}}},$$

where $P_{\mathrm{iT}}$ and $P_{\mathrm{iC}}$ are the risks of the outcome in the treatment group and control groups, respectively, of the $i$th study and $n$ is the number of patients in the respective groups. The weights ($w$) allocated to each of the studies are inversely proportional to the square of the standard error. Thus,

$$w_i = \frac{1}{se_i{}^2},$$

which gives greater weight to those studies with smaller standard errors. The combined effect size is computed by the weighted average as

$$\overline{ES} = \frac{\sum (w_i \times ES)}{\sum w_i},$$

where $\overline{ES}$ is the effect size measure, and it has a standard error given by

$$se_{\overline{ES}} = \sqrt{\frac{1}{\sum w_i}}.$$

Assuming these estimates are distributed normally, the 95% confidence limits are easily obtained by $Lower = \overline{ES} \pm 1.96(se_{\overline{ES}})$.

## The Conventional Approach to Heterogeneity

As can be seen above, the variability within studies is incorporated in the current approach to combining the effects by adjusting based on the variance of the estimates in each individual study. So, if a study reports a higher variance for its RR estimate, it would get lesser weight in the final combined estimate and vice versa. Although statistically very appealing, this approach does not take into account the innate variability that

exists between the studies arising from differences in the study protocols and how well they were executed and conducted. This major limitation has been well recognized and gave rise to the random-effects model approach. Selection of the random-effects model is often carried out using a simple test of homogeneity of the studies involved because of the demonstration that, in the presence of even slight between-study heterogeneity, the fixed-effects model results in inferences that substantially underestimate the variation in the data and in parameter estimates. Nevertheless, despite widespread use of this adjustment for heterogeneity, it is now thought that the choice of fixed-effects or random-effects meta-analysis should not be made on the basis of perceived heterogeneity but rather on the basis of purpose. The fixed-effects meta-analysis tests the null hypothesis that treatments were identical in all trials. When and if this is rejected, then the alternative hypothesis that may be asserted is, "There is at least one trial in which the treatments differed." In other words, the random effects analysis works as a check on the robustness of conclusions from a fixed-effects model to failure in the assumption of homogeneity and cannot go beyond this causal "finding." Because

$$se_{\overline{ES}} = \sqrt{\frac{1}{\sum w_i}},$$

the addition of an external constant will also inflate the variance much more than a redistribution of the weights—if the studies demonstrate varying effects. Furthermore, if a random variable is inserted to inflate the variance based on heterogeneity, it is not clear what aspect of between-trial differences is being assessed and fails to take into account any meaningful differences between the individual studies. Stephen Senn has provided an analytic demonstration of this in his article "Trying to Be Precise About Vagueness." Another problem is that inserting a random variable to inflate the variance based on heterogeneity does not clarify what aspect of between-trial differences is being assessed and fails to take into account any meaningful differences between the individual studies. As such, when used in a meta-analysis of poorly designed studies, it will still result in bad statistics even though there is statistical adjustment for heterogeneity.

## The Quality Effects Approach to Heterogeneity

Suhail A. R. Doi and Lukman Thalib proposed another approach to adjustment for between-study variability by incorporating a *relevant* component (quality) that differs between studies in addition to the weight based on intrastudy differences that is used in any fixed-effects meta-analysis model. The strength of the quality effects meta-analysis is that it allows available methodological evidence to influence subjective random probability and thereby helps to close the damaging gap that has opened up between methodology and statistics in clinical research. This is done by introducing a correction for the quality-adjusted weight of the $i$th study called $\hat{\tau}_i$. This is a composite based on the quality of other studies except the study under consideration and is used to redistribute quality-adjusted weights based on the quality-adjusted weights of other studies. In other words, if Study $i$ is of poor quality, the removed proportion of its quality-adjusted weight is mathematically equally redistributed to all other studies, giving them more weight toward the overall effect size. As studies increase in quality, redistribution becomes progressively less and ceases when all studies are of perfect quality. To do this, weights first have to be adjusted for quality, and one way to incorporate quality scores into such an analysis is as follows:

$$\overline{ES} = \frac{\sum (Q_i \times w_i \times ES_i)}{\sum (Q_i \times w_i)},$$

where $Q_i$ is the judgment of the probability (0 to 1) that Study $i$ is credible, based on the study methodology. The variance of this weighted average is then

$$se^2 = \frac{\sum ({Q_i}^2 \times w_i)}{\left(\sum (Q_i \times w_i)\right)^2}.$$

However, this probabilistic viewpoint on quality-adjusted weights fails to redistribute the removed weight, and Doi and Thalib do this as follows: Given that $\hat{\tau}_i$ is the quality adjustor for the $i$th

study and $N$ is the number of studies in the analysis, then $Q_i$ can be modified to $\hat{Q}_i$ as follows:

$$\hat{Q}_i = Q_i + \left(\frac{\hat{\tau}_i}{w_i}\right),$$

where $\hat{\tau}_i = \left(\sum_{i=1}^{N} \tau_i - \tau_i\right)$ and $\tau_i = \frac{w_i - (w_i \times Q_i)}{N-1}$.

This definition of $\hat{\tau}_i$ has been updated recently to make redistribution proportional to the quality of the other studies rather than equally, as in the definition of $\hat{\tau}_i$ above, and this updated $\hat{\tau}_i$ is given by

$$\hat{\tau}_i = \left(\sum_{i=1}^{N} \tau_i \times N \times \frac{Q_i}{\sum_{i=1}^{N} Q_i}\right) - \tau_i.$$

The final summary estimate is then given by

$$\overline{ES}_{(QE)} = \frac{\sum (\hat{Q}_i \times w_i \times ES_i)}{\sum (\hat{Q}_i \times w_i)},$$

and the variance of this weighted average is then

$$v_{\overline{ES}(QE)} = \frac{\sum (\hat{Q}_i^{\,2} \times w_i)}{\left(\sum (\hat{Q}_i \times w_i)\right)^2}.$$

Although it may seem that $\hat{Q}_i$ is a function of $w_i$, given that $(\hat{Q}_i \times w_i) = (Q_i \times w_i) + \hat{\tau}_i$, it would mean that by multiplying $\hat{Q}_i$ with $w_i$, we are actually adjusting the product of quality and weight by $\hat{\tau}_i$, and by definition, the latter is a function of the quality and weights of other studies excluding this $i$th study. This suggested adjustment has a parallel to the random-effects model, where a constant is generated from the homogeneity statistic

$$Q = \sum (w \times ES^2) - \frac{[\sum (w \times ES)]^2}{\sum w},$$

and using this and the number of studies, $k$, a constant ($\tau^2$) is generated, given by

$$\tau^2 = \frac{Q - (k-1)}{\sum w - \left(\frac{\sum w^2}{\sum w}\right)}.$$

The inverse of the sampling variance plus this constant that represents the variability across the population effects is then used as the weight

$$\left(w_{(R)i} = \frac{1}{se_i^2 + \tau^2}\right).$$

In effect as $\tau^2$ gets bigger, the $se_{\overline{ES}}$ increases, thus widening the confidence interval. The weights, however, become progressively more equal, and in essence, this is the basis for the random-effects model—a form of redistribution of the weights so that outlier studies do not unduly influence the pooled effect size. This is precisely what the quality effects model does too, the only difference being that a method based on quality is used rather than statistical heterogeneity, and $se_{\overline{ES}}$ is not as artificially inflated as in the random-effects model. The random-effects model adds a single constant to the weights of all studies in the meta-analysis based on the statistical heterogeneity of the trials. This method redistributes the quality-adjusted weights of each trial based on the measured quality of the other trials in the meta-analysis.

## Software

This method has been incorporated into version 1.7 of the MIX program, which is a comprehensive free software for meta-analysis of causal research data available from the MIX website.

*Suhail A. R. Doi*

***See also*** Fixed-Effects Models; Meta-Analysis; Random-Effects Models

### Further Readings

Brockwell, S. E., & Gordon, I. R. (2001). A comparison of statistical methods for meta-analysis. *Statistics in Medicine, 20,* 825–840.

Cochran, W. G. (1937). Problems arising in the analysis of a series of similar experiments. *Journal of the Royal Statistical Society, 4,* 102–118.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials, 7,* 177–188.

Doi, S. A., & Thalib, L. (2008). A quality-effects model for meta-analysis. *Epidemiology, 19,* 94–100.

Fleiss, J. L., & Gross, A. J. (1991) Meta-analysis in epidemiology, with special reference to studies of the

association between exposure to environmental tobacco smoke and lung cancer: A critique. *Journal of Clinical Epidemiology, 44,* 127–139.
Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5,* 3–8.
Hardy, R. J., & Thompson, S. G. (1998). Detecting and describing heterogeneity in meta-analysis. *Statistics in Medicine, 17,* 841–856.
Klein, S., Simes, J., & Blackburn, G. L. (1986). Total parenteral nutrition and cancer clinical trials. *Cancer, 58,* 1378–1386.
Senn, S. (2007). Trying to be precise about vagueness. *Statistics in Medicine, 26,* 1417–1430.
Smith, S. J., Caudill, S. P., Steinberg, K. K., & Thacker, S. B. (1995). On combining dose-response data from epidemiological studies by meta-analysis. *Statistics in Medicine, 14,* 531–544.
Tritchler, D. (1999). Modelling study quality in meta-analysis. *Statistics in Medicine, 18,* 2135–2145.
Woolf, B. (1955). On estimating the relation between blood group and disease. *Annals of Human Genetics, 19,* 251–253.

**Websites**

MIX—Meta-Analysis Made Easy: http://www.mix-for-meta-analysis.info

# Quantitative Research

Quantitative research studies produce results that can be used to describe or note numerical changes in measurable characteristics of a population of interest; generalize to other, similar situations; provide explanations of predictions; and explain causal relationships. The fundamental philosophy underlying quantitative research is known as positivism, which is based on the scientific method of research. Measurement is necessary if the scientific method is to be used. The scientific method involves an empirical or theoretical basis for the investigation of populations and samples. Hypotheses must be formulated, and observable and measurable data must be gathered. Appropriate mathematical procedures must be used for the statistical analyses required for hypothesis testing.

Quantitative methods depend on the design of the study (experimental, quasi-experimental, nonexperimental). Study design takes into account all those elements that surround the plan for the investigation, such as research question or problem statement, research objectives, operational definitions, scope of inferences to be made, assumptions and limitations of the study, independent and dependent variables, treatment and controls, instrumentation, systematic data collection actions, statistical analysis, time lines, and reporting procedures. The elements of a research study and experimental, quasi-experimental, and nonexperimental designs are discussed here.

## Elements

### *Problem Statement*

First, an empirical or theoretical basis for the research problem should be established. This basis may emanate from personal experiences or established theory relevant to the study. From this basis, the researcher may formulate a research question or problem statement.

### *Operational Definitions*

Operational definitions describe the meaning of specific terms used in a study. They specify the procedures or operations to be followed in producing or measuring complex constructs that hold different meanings for different people. For example, intelligence may be defined for research purposes by scores on the Stanford-Binet Intelligence Scale.

### *Population and Sample*

Quantitative methods include the target group (population) to which the researcher wishes to generalize and the group from which data are collected (sample). Early in the planning phase, the researcher should determine the scope of inference for results of the study. The scope of inference pertains to populations of interest, procedures used to select the sample(s), method for assigning subjects to groups, and the type of statistical analysis to be conducted.

### *Formulation of Hypotheses*

Complex questions to compare responses of two or more groups or show relationships between

two or more variables are best answered by hypothesis testing. A hypothesis is a statement of the researcher's expectations about a relationship between variables.

### *Hypothesis Testing*

Statements of hypotheses may be written in the alternative or null form. A directional alternative hypothesis states the researcher's predicted direction of change, difference between two or more sample means, or relationship among variables. An example of a directional alternative hypothesis is as follows:

> Third-grade students who use reading comprehension strategies will score higher on the State Achievement Test than their counterparts who do not use reading comprehension strategies.

A nondirectional alternative hypothesis states the researcher's predictions without giving the direction of the difference. For example:

> There will be a difference in the scores on the State Achievement Test between third-grade students who use reading comprehension strategies and those who do not.

Stated in the null form, hypotheses can be tested for statistically significant differences between groups on the dependent variable(s) or statistically significant relationships between and among variables. The null hypothesis uses the form of "no difference" or "no relationship." Following is an example of a null hypothesis:

> There will be no difference in the scores on the State Achievement Test between third-grade students who use reading comprehension strategies and those who do not.

It is important that hypotheses to be tested are stated in the null form because the interpretation of the results of inferential statistics is based on probability. Testing the null hypothesis allows researchers to test whether differences in observed scores are real, or due to chance or error; thus, the null hypothesis can be rejected or retained.

### *Organization and Preparation of Data for Analysis*

Survey forms, inventories, tests, and other data collection instruments returned by participants should be screened prior to the analysis. John Tukey suggested that exploratory data analysis be conducted using graphical techniques such as plots and data summaries in order to take a preliminary look at the data. Exploratory analysis provides insight into the underlying structure of the data. The existence of missing cases, outliers, data entry errors, unexpected or interesting patterns in the data, and whether or not assumptions of the planned analysis are met can be checked with exploratory procedures.

### *Inferential Statistical Tests*

Important considerations for the choice of a statistical test for a particular study are (a) type of research questions to be answered or hypotheses to be tested; (b) number of independent and dependent variables; (c) number of covariates; (d) scale of the measurement instrument(s) (nominal, ordinal, interval, ratio); and (e) type of distribution (normal or non-normal). Examples of statistical procedures commonly used in educational research are *t* test for independent samples, analysis of variance, analysis of covariance, multivariate procedures, Pearson product-moment correlation, Mann–Whitney *U* test, Kruskal–Wallis test, and Friedman's chi-square test.

### *Results and Conclusions*

The level of statistical significance that the researcher sets for a study is closely related to hypothesis testing. This is called the alpha level. It is the level of probability that indicates the maximum risk a researcher is willing to take that observed differences are due to chance. The alpha level may be set at .01, meaning that 1 out of 100 times the results will be due to chance; more commonly, the alpha level is set at .05, meaning that 5 out of 100 times observed results will be due to chance. Alpha levels are often depicted on the normal curve as the critical region, and the researcher must reject the null hypothesis if the data fall into the predetermined critical region. When this occurs, the researcher must conclude that the findings are statistically significant. If the

researcher rejects a true null hypothesis (there is, in fact, no difference between the means), a Type I error has occurred. Essentially, the researcher is saying there is a difference when there is none. On the other hand, if a researcher fails to reject a false null (there is, in fact, a difference), a Type II error has occurred. In this case, the researcher is saying there is no difference when a difference exists. The power in hypothesis testing is the probability of correctly rejecting a false null hypothesis. The cost of committing a Type I or Type II error rests with the consequences of the decisions made as a result of the test. Tests of statistical significance provide information on whether to reject or fail to reject the null hypothesis; however, an effect size ($R^2$, eta$^2$, phi, or Cohen's $d$) should be calculated to identify the strength of the conclusions about differences in means or relationships among variables.

## True Experimental Designs

True experimental designs are the most rigorous quantitative research methods in that they allow the researcher to have full control over the experiment and to assign subjects randomly to groups. Full control and randomization strengthen internal validity; however, external validity may be compromised. Experimental designs are useful for establishing cause-and-effect relationships among variables. One or more variables are systematically manipulated so that effects of the manipulated variables on two or more groups or individuals can be observed. The variable being manipulated is called an independent variable (treatment), and the observed variable is called a dependent variable (measured outcome). True experimental designs satisfy the following criteria: randomization, experimental control(s), experimental treatment(s), hypothesis testing, experimental and control group(s), and standardized research instruments.

### *Criteria*

#### Randomization

Random assignment of subjects to groups helps to control for bias in the selection process. Probability sampling procedures ensure that each participant in a study has an independent and equal chance of being selected for any group.

#### Experimental Control

The researcher should attempt to hold all variables constant that might influence the outcome of the study by allowing only the dependent variable to vary based on participants' responses to the treatment. For example, in an educational study, variables such as achievement motivation, ability level, and satisfaction with school are difficult to control, yet these variables could influence the outcomes of a study. Pretesting, matching, blocking, and using covariates are common controls.

#### Experimental Treatment

One or more interventions or treatments may be manipulated in an experiment. New teaching methods, alternative testing methods, and different curricula are examples of treatments used in educational research.

#### Experimental and Control Groups

An educational researcher could compare reading achievement of two groups of students. The manipulated variable would be method of instruction (whole language or traditional), and the dependent variable would be reading achievement scores. The experimental group is taught reading by the whole-language approach, and the control group is taught reading by the traditional approach. Both groups would be tested after the experiment using the same standardized test. Any change in reading achievement scores may be tentatively attributed to method of instruction.

#### Standardized Test Instruments

A test that meets certain standards or criteria for technical adequacy in construction, administration, and use is said to be standardized. A standardized test provides clear directions for administration and scoring, so that repeated administrations and scoring of the test are carried out systematically. The American Psychological Association, American Educational Research Association, and the National Council on Measurement in Education have established standards that address professional issues and technical characteristics of standardized test development and use in the social sciences.

### *Basic Patterns*

Donald Campbell and Julian Stanley suggested three basic patterns of true experimental research designs: pretest–posttest control group design, posttest-only control group design, and Solomon four-group design.

#### Pretest–Posttest Control Group Design

A pretest is administered to a control group and an experimental group prior to the administration of the treatment. After the experiment, a posttest is administered to both groups, and gain scores from the pretest to the posttest may be compared. Statistically significant differences between gain score means may be computed using a *t* test for independent samples if only two groups are involved.

#### Posttest-Only Control Group Design

When pretesting is not possible or desirable, the posttest-only control group design may be used. Random assignment of subjects to groups serves to ensure equality of groups. Gain scores cannot be computed; otherwise, statistical analysis for the posttest-only design is the same as that for the pretest–posttest control group design.

#### Solomon Four-Group Design

This design is configured so that one experimental group receives the pretest and treatment; one control group receives the pretest, but no treatment; one control group receives the treatment, but no pretest; and one control group receives neither pretest nor treatment. All groups receive the posttest. The Solomon four-group design is a combination of the pretest–posttest control group design and the posttest-only control group design. Essentially, the design requires the conduct of two experiments: one with pretests and one without pretests. The design provides more rigorous control over extraneous variables and greater generalizability of results than either of the previous designs. Because the design does not provide four complete measures for each of the groups, statistical analysis is performed on the posttest scores using a two-way ($2 \times 2$) analysis of variance procedure. Thus, the researcher will be able to ascertain the main effects of the treatment, main effects of pretesting, and the interaction effect of pretesting with the treatment.

## Quasi-Experimental Designs

Quasi-experimental designs are appropriate when random assignment of subjects to groups is not possible. Much of the research in education and psychology is conducted in the field or in classroom settings using intact groups. In such cases, researchers assign treatments randomly to nonrandomly selected subjects. The lack of full control and nonrandom assignment of subjects to groups pose threats to the internal and external validity of quasi-experimental designs. Matching may be used to control for the lack of randomization. In matching, researchers try to select groups that are as similar as possible on all important variables that may affect the outcomes of a study. Pretests are also recommended to control for lack of randomization. Similar scores on a pretest administered to all groups indicate that the groups were matched adequately. There is no doubt that quasi-experiments are weaker than true experiments for making causal inferences; however, information resulting from quasi-experiments is usually better than no information at all.

### *Nonrandomized Control Group, Pretest–Posttest Design*

Using intact groups for data collection is not only more convenient for the researcher, but also more palatable to other stakeholders, for example, educational administrators in educational studies. A random procedure is used to determine which group or groups will be the control and which will be the experimental. In these situations, the researcher cannot assign subjects randomly to groups, so pre-experimental sampling equivalence cannot be ensured. Similar scores on a pretest for the experimental and control groups denote a greater degree of pre-experimental equivalency. After the pretest, the treatment is introduced, and both groups are administered the same posttest. Differences in pre- and posttest mean scores for the control and experimental groups are more credible when pretest means are similar. When results of the pretest reveal that the scores are not similar, more involved statistical procedures are

required to address the measurement error on the pretest.

### *Time Series*

Periodic measures are taken on one group at different intervals over an extended time period. Essentially, a time series design involves a series of pre- and posttest measures. A treatment is introduced, and its effects are assessed based on the stability of the repeated measures and differences from one measure to another.

### *Single-Subject Designs*

The researcher investigates a single behavior, or a limited number of behaviors, of one participant by making initial observations, administering a treatment, and observing and measuring behavior after the treatment to study the effects. Pretreatment observations are necessary to establish baseline data. A simple single-subject design is the AB design, where pretreatment observations and measures (A) are made to establish baseline data, and then treatment (B) is administered and observations and measures (A) are repeated to ascertain changes in the behavior. Monitoring of the treatment usually lasts until it is apparent that the treatment has had some effect on behavior. The sequence of observations and measures (A) and treatment (B) followed by prolonged observations and measures (A) may be repeated, as in the ABA and ABAB designs.

### *Factorial Designs*

Many research questions require the investigation of more than one independent variable. Factorial designs permit the investigation of the effect of one independent variable on the dependent variable while ignoring other independent variables (main effect) and the combined influence (interaction) between two or more independent variables. Notation for factorial designs is usually written to identify the number of levels involved in each independent variable. For example, a "two by three" factorial design ("2 × 3") would have two independent variables (factors) with two levels on the first independent variable and three levels on the second independent variable. All groups receive each level on each of the independent variables. A 2 × 3 design would require six groups. An example of a 2 × 3 factorial design is a study of method of teaching reading (whole language vs. traditional) and motivation (high, medium, low). Three groups would receive whole-language reading instruction and three groups would receive the traditional reading instruction. The dependent variable would be posttest scores on a standardized reading test. Interpretations of the analysis are more complex for factorial designs than for designs discussed previously.

## Nonexperimental Designs

It is not always possible, practical, or ethical to manipulate the research variables of interest. For instance, whether or not parents are involved in their child's education, the teacher's level of job satisfaction, or whether or not a student has transferred from one school to another cannot be manipulated by the researcher. In these cases, the variable has already occurred. For such studies, nonexperimental methods are used. Three common nonexperimental methods of quantitative research are causal comparative (ex post facto), survey, and correlation.

### *Causal Comparative*

Causal comparative studies do not permit the researcher to control for extraneous variables. Changes in the independent variable have already occurred before the study is conducted; thus, these studies are also known as *ex post facto,* meaning "from after the fact." Research questions of interest may be on differences in subjects on dependent variables when subjects have known differences on independent variables, or the extent to which subjects differ on independent variables when they have known differences on the dependent variables.

### *Survey*

Sample surveys, unlike census surveys that query the entire population as in the U.S. Census, collect data from a sample of individuals thought to be representative of the larger (target) population to which the researcher wants to generalize. Surveys are used to collect data over a period of time (longitudinal) or to collect data at one point in time.

Responses are aggregated and summarized, and participant identity should be confidential or anonymous. Statistical procedures for reporting survey data include frequencies, percent, cross-tabulations (cross-tabs), chi-square statistic, phi coefficient, Kendall coefficient, and the gamma statistic.

### *Correlation*

Correlation research is used to explore relationships between or among two or more variables. Correlation studies are useful for establishing predictive validity, establishing test reliability, and describing relationships. Simple correlation procedures involve ascertaining the relationship between two variables, whereas partial correlation procedures are used to control for a variable that may influence the correlation between two other variables. A multiple correlation coefficient (multiple regression) indicates the relationship between the best combination of independent variables and a single dependent variable. Canonical correlation indicates the relationship between a set of independent variables and a set of dependent variables. The kind of correlation coefficient computed depends on the type of measurement scale used and the number of variables.

*Marie Kraska*

***See also*** Data Cleaning; Experimental Design; Hypothesis; Nonexperimental Design; Quasi-Experimental Design; Sample; Survey

### Further Readings

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Myers, R. H. (1990). *Classical and modern regression with applications* (2nd ed.). Pacific Grove, CA: Duxbury.

Ramsey, F. L., & Schafer, D. W. (2002). *The statistical sleuth* (2nd ed.). Pacific Grove, CA: Duxbury.

Salkind, N. J. (2000). *Statistics for people who think they hate statistics*. Thousand Oaks, CA: Sage.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Wiersma, W., & Jurs, S. G. (2005). *Research methods in education: An introduction* (8th ed.). Needham Heights, MA: Allyn & Bacon.

## Quasi-Experimental Design

A scientific experiment is a controlled set of observations aimed at testing whether two or more variables are causally related. William Shadish, Thomas Cook, and Donald Campbell describe two broad types of experiments: (a) randomized experiments, in which study units are randomly assigned to observational conditions; and (b) quasi-experiments, in which study units are not randomly assigned to observational conditions because of ethical or practical constraints. Although it is more difficult to draw causal inferences from quasi-experiments than from randomized experiments, careful planning of quasi-experiments can lead to designs that allow for strong causal inferences.

In order to infer a relationship between cause and effect, three requirements must be met: Cause must precede effect; cause must be related to effect; and, aside from the cause, no alternative explanation for the effect must be plausible. Randomized and quasi-experiments do not differ with respect to the first two requirements. However, with respect to the third requirement, randomized experiments have an advantage over quasi-experiments. Because study units are randomly assigned to conditions in randomized experiments, alternative explanations (e.g., confounding variables) are equally likely across these conditions and can be ruled out. But because quasi-experiments lack random assignment between conditions, alternative explanations are difficult to rule out. This entry focuses on the validity of, common designs of, and inferences drawn from quasi-experiments.

### Validity

Inferences based on an experiment are only as good as the evidence that supports them. The term *validity* is used to refer to the relation between the conclusion of an inference and its supporting evidence. In experimentation, inferences (i.e., conclusions) are valid if they are plausible.

A number of conditions must be met in order to draw a valid inference based on an experiment.

These conditions fall into four categories. First, the *internal validity* of an inference refers to whether the covariation between the experimental manipulation and the experimental outcome does indeed reflect a causal relationship between the manipulation and outcome. Second, *external validity* refers to the generalizability of an inference (i.e., do the results of the experiment apply outside of the experimental setting?). Third, *statistical conclusion validity* refers to the validity of inferences about the covariation between manipulation and outcome. Fourth, *construct validity* refers to the validity of inferences about the higher order construct(s) that the experimental manipulation operationalizes.

### *Threats to Internal Validity*

Factors that influence the nature and strength of inferences are referred to as threats to validity. Of particular relevance to quasi-experimental designs are threats to internal validity as they increase the likelihood that a plausible alternative explanation for the experimental outcome exists. Shadish and colleagues identify the following threats to internal validity:

*Ambiguous temporal precedence:* Lack of clarity about which variable occurred first may yield confusion about which variable is the cause and which is the effect.

*Selection:* Systematic differences over conditions in respondent characteristics that could also cause the observed effect.

*History:* Events occurring concurrently with treatment could cause the observed effect.

*Maturation:* Naturally occurring changes over time could be confused with a treatment effect.

*Regression:* When units are selected for their extreme scores, they will often have less extreme scores on other variables, an occurrence that can be confused with a treatment effect.

*Attrition:* Loss of respondents to treatment or to measurement can produce artificial effects if that loss is systematically correlated with conditions.

*Testing:* Exposure to a test can affect scores on subsequent exposures to that test, an occurrence that can be confused with a treatment effect.

*Instrumentation:* The nature of a measure may change over time or conditions in a way that could be confused with a treatment effect.

*Additive and interactive effects of threats to internal validity:* The impact of a threat can be added to that of another threat or may depend on the level of another threat.

## Common Designs

Because threats to internal validity are prominent in quasi-experiments, care must be taken to ensure that the influence of these threats to validity is minimized. Shadish and colleagues discuss three principles useful in this respect: identification and study of plausible threats to internal validity; design controls that limit threats to internal validity (e.g., control groups, pretest/posttest designs); and specific hypotheses that limit the number of viable alternative explanations (e.g., predicted interactions or inclusion of *nonequivalent dependent variables*, that is, a dependent variable that is predicted not to change because of the manipulation but is expected to respond to threats to internal validity the same way as the dependent variable being studied). Four types of common quasi-experimental designs are discussed, each of which has its own advantages and disadvantages concerning threats to internal validity. This discussion and examples of quasi-experimental designs draw on the work of Shadish and colleagues.

### *Designs Without a Control Group*

#### One-Group Posttest-Only Design

This is a simple design involving a posttest on participants ($O_1$) following a manipulation (X).

$$X \quad O_1$$

For example, suppose knowledge of the causes of sudden infant death syndrome (SIDS) is low and stable within a community. Public health officials create a media campaign to raise awareness regarding factors affecting SIDS. Following the campaign, a sample of citizens is surveyed concerning their knowledge of these factors. If the citizens are aware of factors affecting SIDS ($O_1$), one may infer that this is due to the media campaign (X).

Nonetheless, the one-group posttest-only design is a very weak design. It is impossible to guarantee temporal precedence, and furthermore, nearly all other threats to internal validity may apply. One way to improve this design, although it will remain weak, is to use multiple, unique posttests ($O_{1A}$ through $O_{1N}$).

$$X \quad \{O_{1A} \; O_{1B} \; \ldots \; O_{1N}\}$$

Results of the several posttests can be assessed individually and compared to the hypothesized outcomes for each individual posttest and the manipulation. This decreases the likelihood of an invalid inference based on just a single prediction.

#### One-Group Pretest–Posttest Design

Instead of having a single observation, as in the previous design, the one-group pretest–posttest design has a pretest measure ($O_1$) before manipulation (X) as well as a posttest measure ($O_2$) following treatment.

$$O_1 \quad X \quad O_2$$

For example, Jonathan Duckhart studied the effects of a program to reduce environmental lead in low-income urban housing units in Baltimore. Lead levels in each home were measured at pretest and following the intervention. Lead levels decreased between pretest and posttest, supporting the conclusion that the program was effective.

Because of the pre- and posttest, temporal precedence is more easily established, although the effect could still have been caused by history or maturation. Possible improvements involve using a double pretest,

$$O_1 \quad O_2 \quad X \quad O_3$$

or a nonequivalent dependent variable.

$$\{O_{1A} \quad O_{1B}\} \quad X \quad \{O_{2A} \quad O_{2B}\}$$

A double pretest can give an estimate of biases that may exist in the observed effect of the manipulation at posttest. Differences between the posttest and the second pretest that are similar to differences between the two pretests are likely due to factors other than the manipulation. Using a nonequivalent dependent variable reduces many threats to internal validity because any changes from pretest to posttest not due to the manipulation should affect the original dependent variable as well as the nonequivalent dependent variable.

### *Control Group Designs*

#### Nonequivalent Groups Posttest-Only Design

The nonequivalent groups posttest-only design is similar to the one-group posttest-only design with a control group added (where NR indicates nonrandom assignment).

$$\begin{array}{lll} \text{NR} & \text{X} & \text{O}_1 \\ \hline \text{NR} & & \text{O}_2 \end{array}$$

For example, suppose psychotherapy researchers obtained two samples of individuals with depression: those who have attended psychotherapy and those who have not. If those who attend psychotherapy have fewer depressive symptoms than those who did not, the researchers may conclude that psychotherapy reduced the symptoms.

The added control group is a definite improvement over the design lacking a control group as it indicates which effects occur without the manipulation. However, the design is still rather weak because the experimental and control groups may have differed on many nonmanipulated variables related to outcome. This situation is often referred as *selection bias*. To estimate selection bias, this design can include an independent pretest sample that does not receive the experimental manipulation.

$$\begin{array}{llll} \text{NR} & \text{O}_1 \mid & \text{X} & \text{O}_2 \\ \hline \text{NR} & \text{O}_1 \mid & & \text{O}_2 \end{array}$$

Here, the pretest and posttest samples are independent across time (i.e., they may be sampled at the same moment in time, and thus on different study units).

Further improvements include matching (where study units with similar scores are matched across experimental and control groups), using internal controls (where the control group consists of a sample drawn from a population similar to that of the experimental group), using multiple nonequivalent control groups, and using a predicted interaction (a highly differentiated causal

hypothesis that predicts one particular interaction but excludes others).

#### Nonequivalent Groups Pretest–Posttest Design

The nonequivalent groups pretest–posttest design is similar to the one-group pretest–posttest design with a control group added.

$$\begin{array}{cccc} \text{NR} & \text{O}_1 & \text{X} & \text{O}_2 \\ \hline \text{NR} & \text{O}_1 & & \text{O}_2 \end{array}$$

Note that although similar to the nonequivalent groups posttest-only design with a pretest, the current design includes a dependent pretest, that is, pretest and posttest data are gathered on the same study units. For example, Grace Carter, John Winkler, and Andrea Biddle compared scientists who had received a Research Career Development Award from the National Institutes of Health in order to improve their research careers to those who did not. They found that at posttest, scientists who had received such an award did better than those who did not. However, the former also exceeded the latter at pretest, thus calling into the question the effectiveness of the research awards.

Because pretest and posttest data are gathered on both experimental and control groups, only one of which receives the experimental manipulation, the existence of a possible selection bias may be estimated. Insofar as selection is present, it may magnify the effects of other threats to internal validity (e.g., maturation).

Improvements of this design include a double pretest that allows for the assessment of the selection-maturation threat to internal validity,

$$\begin{array}{ccccc} \text{NR} & \text{O}_1 & \text{O}_2 & \text{X} & \text{O}_3 \\ \hline \text{NR} & \text{O}_1 & \text{O}_2 & & \text{O}_3 \end{array};$$

switching replications, which entails delivering the manipulation to control group at a later date,

$$\begin{array}{ccccc} \text{NR} & \text{O}_1 & \text{X} & \text{O}_2 & \text{O}_3 \\ \hline \text{NR} & \text{O}_1 & & \text{O}_2 & \text{X} \quad \text{O}_3 \end{array};$$

using a reversed treatment control group, in which two manipulations are delivered that are opposite in direction of effect,

$$\begin{array}{cccc} \text{NR} & \text{O}_1 & \text{X}_+ & \text{O}_2 \\ \hline \text{NR} & \text{O}_1 & \text{X}_- & \text{O}_2 \end{array};$$

or direct measurement of threats to validity and incorporating these estimates into statistical analysis of the outcomes.

In addition to these modifications, cohort (successive, comparable groups) controls may also be used to improve the nonequivalent groups pretest-posttest design:

$$\begin{array}{cccc} \text{NR} & \text{O}_1 & & \\ \hline \text{NR} & & \text{X} & \text{O}_2 \end{array}.$$

The first group, which is similar to the second group in relevant aspects, does not receive the experimental manipulation, whereas the second group does. Because of the similarity of the two groups, any differences between them are assumed to be related to the manipulation.

This simple cohort control design may be further improved by adding pretests.

$$\begin{array}{ccccc} \text{NR} & \text{O}_1 & \text{O}_2 & & \\ \hline \text{NR} & & & \text{O}_3 \quad \text{X} & \text{O}_4 \end{array}$$

#### *Interrupted Time-Series Designs*

In an interrupted time series design, the same variable is measured repeatedly over time.

$$\text{O}_1 \ \text{O}_2 \ \text{O}_3 \ \text{O}_4 \ \text{O}_5 \ \text{X} \ \text{O}_6 \ \text{O}_7 \ \text{O}_8 \ \text{O}_9 \ \text{O}_{10}$$

A change in intercept or slope of the time series is expected at the point in time where the manipulation was delivered. For example, A. John McSweeny studied the effects of the Cincinnati Bell phone company instituting a charge of 20 cents per call for local directory assistance. At the point in time when this charge was added, a significant and immediate drop in the number of local directory assistance calls is visible in the data. This illustrates how a change in intercept of a time series can be used to judge the effect of a manipulation. A change in slope can also be used to judge the effect of a manipulation. For example, Julian Roberts and Robert Geboyts studied the impact of the reform of Canadian sexual assault law instituted in order to increase reporting of sexual assault crimes. They found a relatively flat slope in the years before the reform (that is, the reported number of

assaults remained fairly constant), whereas in the years following the reform, a steady increase in the reported number of assaults was observed.

A major threat to this type of design is history—that is, other factors occurring at the same time as the manipulation that may cause the outcome under investigation. Several improvements can be made to this design. A nonequivalent control group may be added:

$$\begin{array}{ccccccccccc} O_1 & O_2 & O_3 & O_4 & O_5 & X & O_6 & O_7 & O_8 & O_9 & O_{10} \\ \hline O_1 & O_2 & O_3 & O_4 & O_5 & & O_6 & O_7 & O_8 & O_9 & O_{10} \end{array}$$

A nonequivalent dependent variable may be added:

$$\begin{array}{ccccccccccc} O_{A1} & O_{A2} & O_{A3} & O_{A4} & O_{A5} & X & O_{A6} & O_{A7} & O_{A8} & O_{A9} & O_{A10} \\ \hline O_{B1} & O_{B2} & O_{A3} & O_{B4} & O_{B5} & & O_{B6} & O_{B7} & O_{B8} & O_{A9} & O_{A10} \end{array}$$

The manipulation may be implemented and subsequently removed at a known time:

$$O_1\ O_2\ O_3\ O_4\ O_5\ X\ O_6\ O_7\ O_8\ O_9\ \not{X}\ O_{10}\ O_{11}\ O_{12}\ O_{13}$$

This may also be done multiple times, resulting in a multiple replications design:

$$O_1\ O_2\ X\ O_3\ O_4\ \not{X}\ O_5\ O_6\ X\ O_7\ O_8\ \not{X}\ O_9\ O_{10}\ X\ O_{11}\ O_{12}\ \not{X}\ O_{13}\ O_{14}$$

Delivering the manipulation to two nonequivalent groups at different times results in a switching replications design:

$$\begin{array}{ccccccccccc} O_1 & O_2 & O_3 & X & O_4 & O_5 & O_6 & O_7 & O_8 & O_9 & O_{10} & O_{11} \\ \hline O_1 & O_2 & O_3 & & O_4 & O_5 & O_6 & O_7 & O_8 & X & O_9 & O_{10} & O_{11} \end{array}$$

### *Regression Discontinuity Design*

In regression discontinuity designs, experimental manipulation is based on a cutoff score ($C$) on an assignment variable ($O_A$) measured before manipulation.

$$\begin{array}{cccc} O_A & C & X & O_2 \\ O_A & C & & O_2 \end{array}$$

If the manipulation has an effect, then regression analysis of the data obtained from this design should reveal a discontinuity at the cutoff score corresponding to the size of the manipulation effect. If the manipulation has no effect, the regression line should be continuous. For example, Charles Wilder studied how the institution of the Medicaid program affected medical visits and found, perhaps unsurprisingly, that household income was positively correlated with medical visits. More importantly, however, the data also showed a dramatic increase in medical visits at the cutoff score for Medicaid eligibility, which supports the inference that the Medicaid program does indeed stimulate medical visits.

Differential attrition may cause a discontinuity in the regression line that resembles an effect of the manipulation and thus is a threat to internal validity. History is a plausible threat to validity if factors affecting the outcome occur only for study units on one side of the cutoff.

## Inferences

Randomized experiments allow for solid inferences about a proposed causal relation of two variables. However, randomization is often not practically or ethically possible, making a quasi-experimental design necessary. Although threats to internal validity are often highly plausible in quasi-experimental designs, researchers can still draw valid causal inferences if they identify plausible threats to internal validity and select quasi-experimental designs that address those threats. Often, the design adjustments require elaborate changes, such as administering and removing the intervention multiple times. However, a small, simple change often can make a large difference, such as adding a nonequivalent dependent variable. In either case, causal inferences are strengthened, which is the primary purpose of experimentation.

*Scott Baldwin and Arjan Berkeljon*

***See also*** Cause and Effect; Experimental Design; Internal Validity; Research Design Principles; Threats to Validity; Validity of Research Conclusions

### Further Readings

Carter, G. M., Winkler, J. D., & Biddle, A. K. (1987). *An evaluation of the NIH research career development award*. Santa Monica, CA: RAND.

Duckhart, J. P. (1998). An evaluation of the Baltimore Community Lead Education and Reduction Corps

(CLEARCorps) program. *Evaluation Review, 22,* 373–402.
McSweeney, A. J. (1978). The effects of response cost on the behavior of a million persons: Charging for directory assistance in Cincinnati. *Journal of Applied Behavioral Analysis, 11,* 47–51.
Roberts, J. V., & Gebotys, R. J. (1992). Reforming rape laws: Effects of legislative change in Canada. *Law and Human Behavior, 16,* 555–573.
Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin.
Wilder, C. S. (1972). *Physician visits, volume, and interval since last visit, U.S. 1969* (No. Series 10, No. 75; DHEW Pub. No. HSM 72–1064). Rockville, MD: National Center for Health Statistics.

# Quetelet's Index

Quetelet's Index, more commonly known as the body mass index (BMI), is a measure of weight relative to height. Originally developed by 19th-century mathematician Lambert Adolphe Jacques Quetelet, it is a standard measurement that is widely used by health professionals and researchers as an index of body fat. The formula used to calculate this index is weight (in kilograms) divided by height (in meters squared). Height in meters can be calculated by dividing height in centimeters by 100. For example, a person who weighs 71 kilograms and is 165 centimeters tall has a BMI of 26 [$(71\ \text{kg})/(1.65\ \text{m})^2 = 26$]. Alternatively, BMI can be calculated by dividing weight (in pounds) by height (in inches squared) multiplied by a conversion factor of 703. For example, a person who weighs 145 pounds and is 65 inches (or 5′ 5″) tall has a BMI of 24 {$[(145\ \text{lbs.})/(65\ \text{in.})^2] \times 703 = 24$}.

Although the body mass index is not a direct measure of adiposity (body fat), it is correlated with direct measures of body fat, such as dual energy x-ray absorptiometry and underwater (hydrostatic) weighing. The advantages of using BMI as a proxy for adiposity are that it is inexpensive, it is easy to obtain, it does not require extensive training, and it is simple to calculate relative to other methods.

## Interpretation for Adults

BMI is often used to classify overweight and obesity, but can be interpreted for the entire weight spectrum. Typically, a person is considered underweight if he or she has a BMI of 18.5 or less and normal weight between 18.5 and 24.9. A BMI of 25.0 to 29.9 is classified as overweight, and 30 or above is considered obese. The obese category can be further subdivided into Class I obesity (30.0 to 34.9), Class II obesity (35.0 to 39.9), and Class III or extreme obesity (40 or above). Recent estimates suggest that two thirds of the U.S. adult population are classified as overweight according to these categories, and nearly one third are obese.

## Interpretation for Children and Adolescents

Although the calculation for BMI is the same for adults and children, the interpretation differs. For adults (aged 20 years and older), BMI is interpreted the same for both men and women using the categories listed above. However, for children and adolescents (aged 2 through 19 years), the interpretation is based on age- and sex-specific percentiles, which reflects the fact that adiposity changes with age and differs among boys and girls. Therefore, it is not appropriate to use the BMI categories for adults to interpret BMI and determine the weight category for children and adolescents. BMI can be used to identify children and adolescents who are either currently overweight or at risk of becoming overweight, based on BMI-for-age growth charts provided by the Centers for Disease Control and Prevention. After obtaining the BMI using the method described above, this number is plotted on the BMI-for-age growth charts for the appropriate sex to yield a percentile ranking, which indicates how the child's BMI compares to children of the same age and sex. Although children are usually not classified as obese, the current recommendation is that BMI values that meet or exceed the 95th percentile of the BMI growth charts for their age and sex should be categorized as overweight, and those who are in the 85th percentile to the 95th percentile are classified as at risk of overweight. These growth curves are J-shaped and were constructed to identify BMI scores that show a trajectory toward overweight ($\text{BMI} \geq 25$) or obesity ($\text{BMI} \geq 30$) in adulthood.

Childhood overweight is of particular concern because research suggests that it frequently tracks into adulthood, with the vast majority of obese adolescents going on to become obese adults.

## Applications

BMI is often used to identify individuals who are at risk for developing weight-related diseases, such as hypertension, high cholesterol, Type 2 diabetes, and coronary heart disease, with increased risk as BMI class increases. Epidemiological research indicates that a BMI equal to or greater than 30 (in adults) is strongly correlated with both morbidity and mortality. Therefore, BMI provides a systematic and efficient method of evaluating risk for adverse health consequences, identifying individuals who may benefit from weight management interventions, as well as tracking changes in body mass over time. However, it should be noted that the BMI categories of overweight and obese identify *relative risk,* rather than *absolute risk,* and therefore must take into consideration other existing risk factors for disease. A high BMI alone does not necessarily indicate health risk; therefore, additional assessments such as family history, dietary and physical activity habits, waist circumference, cholesterol level, and blood pressure are necessary to determine risk status.

## Reliability

Generally, BMI has been found to be a reliable indicator of body fatness, with a moderate to strong correlation. However, some have argued that although BMI is a valuable tool, it does have limitations that must also be taken into consideration. For example, BMI does not distinguish between lean and fat body mass, nor does it take into account the regional distribution of body fat (central vs. peripheral adiposity), both of which are important factors in assessing disease risk. Additionally, one assumption of the BMI is that adiposity can be adequately represented by adjusting weight for height regardless of age, race, or sex. Yet there is evidence that this association varies among different populations. For example, it may overestimate body fat in athletes or those with a muscular build, and may be misleading for certain ethnic groups. Moreover, an increase in the ratio of fat to lean body mass is observed as individuals age, even among those who maintain a consistent BMI. Although BMI remains a widely used and useful tool, these considerations are important because the health risks associated with overweight and obesity are related to excess body fat, rather than weight per se.

*Katherine Presnell*

***See also*** Percentile Rank; Reliability

### Further Readings

Aronne, L. J., & Segal, K. R. (2002). Adiposity and fat distribution outcome measures: Assessment and clinical implications. *Obesity Research, 10,* 14S–21S.

Centers for Disease Control and Prevention. (n.d.). *National Health and Nutrition Examination Survey: 2000 CDC growth charts: United States*. Retrieved from http://www.cdc.gov/growthcharts

Cole, T. J., Bellizzi, M. C., Flegal, K. M., & Dietz, W. H. (2000). Establishing a standard definition for child overweight and obesity worldwide: International survey. *British Medical Journal, 320,* 1240–1243.

Deurenberg, P., Yap, M., & van Staveren, W. A. (1998). Body mass index and percent body fat: A meta analysis among different ethnic groups. *International Journal of Obesity, 22,* 1164–1171.

Gallagher, D., Visser, M., Sepulveda, D., Pierson, R. N., Harris, T., & Heymsfield, S. B. (1997). How useful is body mass index for comparison of body fatness across age, sex, and ethnic groups? *American Journal of Epidemiology, 143,* 228–239.

Garrow, J. S., & Webster, J. (1985). Quetelet's index (W/H2) as a measure of fatness. *International Journal of Obesity, 9,* 147–153.

Hedley, A. A., Ogden, C. L., Johnson, C. L., Carroll, M. D., Curtin, L. R., & Flegal, K. M. (2004). Overweight and obesity among US children, adolescents and adults, 1999–2002. *Journal of the American Medical Association, 291,* 2847–2850.

National Institutes of Health/NHLBI. (1998). Clinical guidelines on the identification, evaluation, and treatment of overweight and obesity in adults: The evidence report. *Obesity Research, 6,* 51S–209S.

Ogden, C. L., Carroll, M. D., Curtin, L. R., McDowell, M. A., Tabak, C. J., & Flegal, K. M. (2006). Prevalence of overweight and obesity in the United States, 1999–2004. *Journal of the American Medical Association, 295,* 1549–1555.

Prentice, A. M., & Jebb, S. A. (2001). Beyond body mass index. *Obesity Reviews, 2*(3), 141–147.

# Quota Sampling

The target population in a sampling survey can often be divided into several subpopulations based on certain characteristics of the population units. Some common variables used to categorize population include gender, age, location, education level, and so forth. It is also common in a sampling survey situation that investigators might require certain propositions or numbers for each type of sampling unit included in the sample for various reasons. For instance, a *representative sample* is often considered one of the most important criteria when evaluating the quality of a sampling survey, and a sample is considered to be representative if the sample structure is the same or close to the population structure; if one is to conduct a survey where gender is considered an influential factor of the response, the investigator needs to draw a sample where the proportion of each gender is as close to the target population as possible. Nevertheless, there are also cases when investigators' subjective judgments play an important role, for example, in a marketing survey of the acceptance level of a new perfume product, the investigator may likely consider a sample with a majority of female interviewees over males appropriate. Bear in mind, other motives or objectives are also possible.

All that being said, in reality, it has never been an easy task to randomly select a sample where the number of units belonging to each subpopulation are fixed; yet many sampling designs developed in the past were targeted toward that end. Several of these designs include stratified sampling designs, proportional allocation of sample sizes, proportional sampling, and quota sampling. Among all, quota sampling is one of the most commonly used sampling selection methods in opinion and marketing survey. The sampling method could be viewed as a different class of sampling designs that shares similar sampling principles. The practical popularity portioned to quota sampling is due mainly to its ability to provide samples with the desired numbers or proportions of each subpopulation at a limited sampling cost. Such an advantage is often compromised by the lack of legitimate statistical inference, as often criticized by statisticians. In the following sections, the general sampling principle of quota sampling and the usual methods used to construct the inference for the population quantity of interest, along with a short review of the advantages and disadvantages, are discussed. A possible manner to reach a balanced solution between sampling convenience and proper inference is also introduced.

## Principle of Quota Sampling

In quota sampling, the quotas of each type of sampling unit are determined in advance, whereby investigators would look to fill in each quota. The quotas are fixed numbers for each type of unit that investigators would like to include in the sample. For instance, an investigator would like to have a sample in which the proportions of each gender by different age levels are the same as the population, and the sample size is 1,000. Suppose that the population proportion is given as shown in Table 1.

The *quota* of each type of sampling unit is the sample size multiplied by the population proportion, and the results are shown in Table 2.

Up until now, quota sampling appears very similar to stratified sampling (i.e., the proportional allocation of sample size emanated). If the within-subpopulation samples were selected by some probability sampling design, the quota sampling in this case can be viewed the same as stratified sampling. However, there is no specific rule as to how the quotas should be filled in quota sampling. As mentioned previously, quotas are sometimes

**Table 1** Population Proportions

| | *Age* | | | | |
|---|---|---|---|---|---|
| *Gender* | *20–29* | *30–39* | *40–49* | *50–59* | *60 and Above* |
| Female | 0.10 | 0.10 | 0.09 | 0.11 | 0.12 |
| Male | 0.10 | 0.08 | 0.1 | 0.11 | 0.09 |

**Table 2** Quotas

| | *Age* | | | | |
|---|---|---|---|---|---|
| *Gender* | *20–29* | *30–39* | *40–49* | *50–59* | *60 and Above* |
| Female | 100 | 100 | 90 | 110 | 120 |
| Male | 100 | 80 | 100 | 110 | 90 |

determined based on subjective judgments, or whatever the investigators consider appropriate; as long as the sampling method is one with a fixed number of different types of sampling, it can be considered a type of quota sampling. Having said that, the term *quota sampling* is often used when the within-subpopulation sample is selected by some relatively nonrandomized designs, such as certain judgmental sampling or convenient sampling designs.

## Inference

Consider a population with population size $N$ is partitioned into $H$ subpopulations, and the sizes of each subpopulation are $N_h$, $h = 1, \ldots, H$ such that

$$N = \sum_{h=1}^{H} N_h.$$

Let $y_{hi}$ be the value of the population variable of interest associated with the $i$th units in the $h$th subpopulation. Usually, the population quantities of primary interest in a survey sampling case would include the overall population total $\tau$, population mean $\mu$, subpopulation totals $\tau_h$, and subpopulation mean $\mu_h$. The population proportion is a special case of the population mean. The definition of the overall population $\tau$ is

$$\tau = \sum_{h=1}^{H} \sum_{i=1}^{N_h} y_{hi} = \sum_{h=1}^{H} \tau_h, \tag{1}$$

where $\tau_h = \sum_{i=1}^{N_h} y_{hi}$. The population mean $\mu$ is

$$\mu = \frac{1}{N}\tau = \frac{1}{N}\sum_{h=1}^{H} \tau_h = \sum_{h=1}^{H} \frac{N_h}{N}\mu_h, \tag{2}$$

where $\mu_h = \tau_h/N_h$. The population variance of the $h$th subpopulation is defined as

$$\sigma_h^2 = \frac{1}{N_h - 1}\sum_{h=1}^{H} (y_{hi} - \mu_h)^2 \tag{3}$$

In addition, let $s_h$ be the with-subpopulation sample of the $h$th subpopulation, that is, $s_h$ is an index set that is a subset of $\{1, 2, \ldots, N_h\}$ and contains the labels of units selected. The sample mean $\overline{y}_h$ and the sample variance $S_h^2$ for the within-subpopulation samples are

$$\overline{y}_h = \frac{1}{n_h}\sum_{i \in s_h} y_{hi} \tag{4}$$

and

$$S_h^2 = \frac{1}{n_h - 1}\sum_{i \in s_h} (y_{hi} - \overline{y}_h)^2, \tag{5}$$

respectively. $n_h$ is the within-subpopulation sample size (quota) of the $h$th subpopulation and $\sum_{h=1}^{H} n_h = n$, the total sample size.

Because there is no specific regulation regarding how the quotas are filled in quota sampling, the inferences of the subpopulation and population totals/means vary depending on the with-subpopulation sampling selection. Usually, $\overline{y}_h$ and $N_h\overline{y}_h$ would be used to estimate $\mu_h$ and $\tau_h$, respectively. Consequently,

$$\hat{\tau} = \sum_{h=1}^{H} N\overline{y}_h \tag{6}$$

and

$$\hat{\mu} = \frac{1}{N}\sum_{h=1}^{H} N\overline{y}_h \tag{7}$$

would be used to estimate the overall population total $\tau$ and mean $\mu$.

If the quota sampling proceeds as the stratified random sampling, that is, the quotas are filled with within-subpopulation simple random sampling without replacement design and the selections are independent from one subpopulation to another, then Equations 6 and 7 are unbiased estimators and the associated confidence intervals based on Central Limit Theorem can be constructed accordingly. If the quotas are filled as general stratified sampling, the associated inferences of $\tau$, $\mu$, $\tau_h$, and $\mu_h$ can be constructed as described in Steven K. Thompson's *Sampling*.

Following the above, the quotas are usually filled with certain nonrandomized designs for sampling convenience; for example, interviewers tend to interview qualified interviewees who are more readily accessible. That said, Equations 6 and 7 can be applied to estimate $\tau$ and $\mu$. Under such

circumstances, it is vital to remember that these estimators are no longer design-based. In fact, one would not be able to evaluate the property or performance of these estimators from a design-based perspective because one would lose the ability to appreciate the accuracy and precision of the estimators by bias and mean-squared error. One possible solution to establish the design-based inference under quota sampling is advised by Mohammad Salehi and Kuang-Cho Chang. Salehi and Chang describe a multiple inverse sampling scheme and the associated inferences to compromise the sampling convenience and a legitimate inference. If a quota sampling can proceed according to the principle of inverse sampling instead of a judgmental design, it is possible to obtain a reasonable statistical inference.

On the other hand, the lack of design-based inference is often considered a major drawback of quota sampling; different authors have thus suggested a model-based inference. With such an approach, the population vector $\mathbf{y} = (y_1, \ldots, y_N)$ is regarded as a constant vector in the design-based approach. The distribution of $\mathbf{Y}$ can be specified by a density function $f(\mathbf{y}; \theta)$, and the randomness used to construct the related inference will be introduced accordingly with the sampling design. Nevertheless, the employment of a model-based approach does not mean that the sampling design is always condonable; such degree depends on the types of inference method and the manner in which the sample is selected.

There are usually two approaches with $t$ model-based inferences—sample-based and data-based. The most well-known example of a sample-based inference, where the inference depends on the conditional distribution given the sample, could be the best linear unbiased estimator, usually referred as BLUE. On the other hand, with the data-based approach, the inference depends on the conditional distribution given the data, such as the likelihood-based inference. If the sampling units are selected independently from $\mathbf{y}$ and other unknown parameters, the sampling design in general can be ignored from both the sample-based and data-based inferences. If the sampling selection $\mathbf{y}$ depends solely on observations, the design can be ignored from the data-based inference, but not the sample-based inference. By the same token, if the sampling selection depends on the unobserved part of $\mathbf{y}$ and/or other unknown parameters, then even the model-based inference needs to take the randomness as introduced by the sampling design into account. Nevertheless, if the quota sample is selected based on subjective judgment, it could be problematic to ignore the design regardless of whether a sample-based or data-based inference is used.

## Advantages and Disadvantages

Quota sampling appears to be a reasonable sampling selection method because of its ability to provide a sample with structure as the investigator would ask, and at a lower sampling cost. As previously mentioned, investigators usually prefer to have a representative sample similar to the population structure; thus, quota sampling seems like a natural choice to draw such samples in a cost-efficient manner. What is often overlooked, nonetheless, is the true representation of the samples, and quite frequently, their representing intensity is low. This can be due to the investigator's inability to include all the influential factors when he or she decides on the factors to categorize the population. For instance, the usual factors used to stratify the population for a political opinion survey might include gender, age, race, and location. However, depending on different issues of interest, other factors can play an important role as well, such as socioeconomic status and educational level. Hence, a sample that appears to be representative could be biased in some ways. Furthermore, with the usual judgment or convenient sampling selection that is used to fill in the quotas, it is known that the interviewees who are easy (or difficult) to reach have certain systematic patterns, which explains the existence of significant yet immeasurable selection biases. Having said that, quota sampling still has its considerable advantage of less sampling effort, which sometimes might be the prime criterion in a survey sampling case. Moreover, it is possible that quota sampling can outperform probability sampling. It is also viewed as a useful sampling design for marketing research. Rather than the possible poor outcome, the real disadvantage of quota sampling points to the challenges faced when trying to evaluate its performance. The

model-based inference might be a possible solution. However, as already discussed, usage of model-based inference does not mean that the design can always be ignored. Additionally, one would always need to consider the validity of the assumed model. The inverse sampling strategy described in Salehi and Chang is another potential solution to use quota sampling with a legitimate probability sampling scheme. Nevertheless, it is recommended that one should be cautious with the inference based on the data drawn by quota sampling.

*Chang-Tai Chao*

***See also*** Convenience Sampling; Sample Size; Stratified Sampling

**Further Readings**

Cochran, W. G. (1977). *Sampling techniques*. New York: Wiley.
Deville, J. (1991). A theory of quota surveys. *Survey Methodology, 17,* 163–181.
Lohr, S. L. (1999). *Sampling: Design and analysis*. N. Scituate, MA: Duxbury.
Salehi, M. A., & Chang, K.C. (2005). Multiple inverse sampling in post-stratification with subpopulation sizes unknown: A solution for quota sampling. *Journal of Statistical Planning and Inference, 131,* 379–392.
Scheaffer, R. L., Mendenhall, W., & Ott, L. (2006). *Elementary survey sampling*. N. Scituate, MA: Duxbury.
Stephan, F., & McCarthy, P. J. (1958). *Sampling opinions*. New York: Wiley.
Thompson, S. K. (2002). *Sampling*. New York: Wiley.
Thompson, S. K., & Seber, G. A. F. (1996). *Adaptive sampling*. New York: Wiley.

# R

## R

The statistical language R has proven to be a popular choice among statisticians, researchers, and organizations for statistical analysis and graphs. This is partly due to the free nature of the program and the power of its programming capabilities. Other strengths include its exceptionally good support, its continual improvement (updates occur almost monthly at times), and its very active user community. R is a model example of the benefits of free and open-source software, and a wealth of contributed documentation is freely available. In recent years, the R language has amassed a growing, and very strong, supporting literature, which consists of several introductory texts and companions to existing textbooks and which implements modern statistical methods, regression modeling strategies, and specialized types of models. R integrates very well with LaTeX and also has very powerful graphics capabilities. For some, the lack of a powerful graphical user interface (i.e., a point-and-click operation) is one of R's most glaring weaknesses.

Presently, R is an open-source software for statistical computing and graphics that can be freely downloaded from the R-project website and runs on several operating systems, such as Mac OS X, Microsoft Windows, and several UNIX platforms. This entry describes the history, general characteristics, and strengths and weaknesses of R and compares R with other statistical software.

### History

In the early 1990s two professors, Ross Ihaka and Robert Gentleman, at the University of Auckland became interested in developing a statistical software that could be used to teach their introductory data analysis courses and that ran efficiently in their teaching laboratories. Since both authors were familiar with the S language, created in the Bell Laboratories by Rick Becker, John Chambers, and Allan Wilks, R was developed as an implementation of S. It gained its name partially as a play on the name S and partially due to the authors' first names starting with the letter R. An initial version of R was made available to the public in 1993 at StatLib, a system for electronically distributing statistical software. After much constructive feedback and encouragement, Ihaka and Gentleman made the source code available in 1995. Then 2 years later, as a result of the large volume of suggestions and contributions overwhelming their capacity to make changes to the source code and report back to the community, a core group of developers was established to make changes to the source code, and The R Project was born. As The R Project grew, its goals shifted toward optimizing and expanding R's capabilities while continuing to provide support and maintenance for the software.

### General Characteristics

R is a command-based environment that supports a wide variety of statistical and mathematical

operations. At its simplest function level, R can be used as a calculator, and at higher levels it can execute complex simulations and statistical analyses.

### *Data Type and Objects*

The items that R operates on include vectors and matrices of numerical or logical values and vectors of character strings. Such items can be assigned to objects that can be saved into files and used in later functions. During any session, the names of all objects in use (the workspace) can be accessed with the command: `objects()`

Objects can be removed from the workspace using the command: `rm(`*`object name`*`)`

Whether objects are saved during a session or not, the user is prompted to save the workspace before exiting the R session.

### *Data Entry*

There are two main ways of entering data into R: manually by assigning values to vectors using the function "`x <- c(a, b, c, d)`" (where a, b, c, and d correspond to the values assigned to a vector) and automatically by reading a data set into R using the `read.table()` command. (This command will open a window and allow the user to choose a data file to be read.) However, the data set being imported has to contain the following characteristics:

- The first row of the data set should have the names of each variable being imported.
- The first column should contain the row labels.

If the data are entered manually into R, the command `newname <- edit(oldname)` allows editing and assigns the corrected data frame to a new object. However, for reading data sets into R, it is easier to clean and format the data in other programs, such as Microsoft Excel or SPSS, an IBM company formerly called PASW (R) Statistics, and then import the file.

When values are missing, R assigns a special value of NA, standing for "not available," to such locations in the object. In general, any operation on an NA becomes an NA, thus indicating that the result of the operation cannot be known and therefore is not available.

R users can construct data sets in programs such as Excel or Notepad and then bring those files into R for further analysis. Data input into R are handled through a series of options, including the following:

- `read.table`
- `read.csv`
- `read.csv2`
- `read.delim`
- `read.delim2`

The primary information that is required to bring the data into R is the location of the file (e.g., a folder on the user's hard drive or elsewhere) and information regarding how the data are formatted. An easier way to import data without typing out the file location is with the `file.choose( )` command. This command will bring up a window in R that will allow the selection of a file by the use of a mouse. For example, `read.table (file.choose())` will import a text file and prompt the user to locate the file in a window.

### *Data Analysis and Packages*

Several statistical analysis methods have been implemented in R and are included as part of the R standard packages, and methods not present in the standard packages can be obtained from the Comprehensive R Archive Network (CRAN) family of Internet sites. In general, packages on the CRAN website consist of functions written by users to accommodate their analysis needs. However, on some occasions, such expansion packages give access to less common statistical methods, such as the BRugs package for Bayesian analysis. Once a package is downloaded from the CRAN site, it needs to be loaded into the R session before it can be accessed. This is accomplished with the following command: `library(`*`package`*`)`.

R provides a comprehensive set of statistical tables and functions to evaluate cumulative distribution and probability density functions and to perform simulations based on identified distributions. For example, the command `unif(1,5)` corresponds to a uniform distribution with lower boundary of 1 and upper boundary of 5. Following are a few examples of the command and the arguments to be specified for each distribution:

| *Distribution* | *Command* |
|---|---|
| binomial | `binom(size, prob)` |
| normal | `norm(mean, sd)` |
| uniform | `unif(min, max)` |
| chi-square | `chisq(df, ncp)` |
| *F* | `f(df1, df2, ncp)` |
| Poisson | `pois(lambda)` |
| Student's *t* | `t(df, ncp)` |

The statistical analysis functions built into R include linear and nonlinear modeling, statistical tests, classification, and mixed models.

### *Graphical Facilities*

R has a very powerful graphing environment that supports a large variety of data representation. The graphing facility allows the display of common statistical graphs as well as the construction of more complex visual representations. There are three classes of plotting commands: high level, low level, and interactive.

High-level commands are responsible for creating new plots, given the provided data. The plot titles, axes, and labels are generated automatically and can be changed by the user in the command lines. The most commonly used graphing function is the `plot(x,y)` command. In this case, the type of graph being displayed depends on the nature of the arguments (*x* and *y*). For example, if *x* and *y* are numeric vectors, the plot command will produce a scatterplot of *y* against *x*. In addition, R supports the plotting of matrices of data. For example, if *x* is a numeric matrix or database, the command `pairs(x)` will plot every column of *x* against every other column of *x* and will arrange all resulting plots in a matrix with plot scales constant over the rows and columns of the matrix. Several other common graphs can be created by high-level commands; the following are a few examples of commands and their graphs:

`hist(argument)` : creates a histogram of the data

`boxplot(argument)` : creates a boxplot

`matplot(argument)` : plots several variables from a table of data.

Low-level commands are used to add more information to a plot, such as extra points, lines, and labels. For example, the command `point(x, y)` adds points to an existing plot. The following are some of the most common low-level graphing commands:

| *Command(s)* | *Function* |
|---|---|
| `lines, abline` | Add lines |
| `points` | Adds points |
| `text, mtext` | Add text |
| `legend` | Adds a legend |
| `axis` | Adds axes |
| `box` | Adds box around all axes |
| `par` | Plotting parameters |
| `colors, palette` | Introduce colors |

Interactive commands allow users to remove or add information to a plot by using a mouse to select points in the plot. The simplest interactive function is the `locator()` command, which lists the *x* and *y* positions of the selected points.

Finally, graphs and visual displays beyond the basic level can be created with packages available at the CRAN site, such as Grid and Lattice.

### *Help*

R is equipped with a help function that one can use easily by inquiring the program about the argument. This can be done in four main ways:

`help(`*argument*`)`

`?` *argument*

`?` *help*

`help.start(`*argument*`)`

However, for special characters, the argument should be included between double quotes (for example: `help(''*'')`. In addition, the `help.search` command allows a search through available help topics, whereas the command `example(`*topic*`)` will bring up examples related to the typed topic.

## Comparisons to SAS and SPSS

Two of the most popular commercial statistical software packages are SAS and SPSS. SAS and SPSS both use one main data structure, the data set. In contrast, R allows for many different types of data

structures (the *data frame* is the R equivalent of what users of SAS and SPSS think of as data sets). SAS and SPSS data sets are always viewed as a rectangle with *variables* in the columns and *records* in the rows. SAS calls these records *observations,* and SPSS calls them *cases*. R documentation uses *variables* and *columns* interchangeably. It usually refers to observations or cases as *rows*.

R data frames have a formal place for an ID variable that R calls *row labels*. SAS and SPSS users typically have an ID variable containing an observation or case number or perhaps a subject's name. But this variable is like any other unless the user runs a procedure that identifies observations. A user can use R this way too, but procedures that identify observations may do so automatically if the user sets his or her ID variable to be official row labels. When a user does that, the variable's original name (ID, subject, Social Security number) vanishes. The information is used automatically when it is needed. Another data structure R uses frequently is the vector. A vector is a single-dimensional collection of numbers (*numeric vector*) or character values (*character vector*) such as variable names.

Variable names in R may be any length and may consist of letters, numbers, or the period and should begin with a letter. Underscores are not allowed, so `my_data` is not a valid name, but `my.data` is. However, if the user always puts quotes around a variable (object) name, the name can be any nonempty string. Unlike SAS, the period has no meaning in the name of a data set in R. Case matters, so a user can have two variables, one named `myvar` and another named `MyVar`, in the same data frame, although that is not recommended. Some add-on packages tweak names; for example, "`Save`" represents a compatible, but enhanced, version of a built-in function like the lowercase "`save`."

R has several operators that are different from SAS or SPSS. The assignment operator for R has traditionally not been the equal sign (although it will work) but rather the two symbols "`<-`" typed right next to each other. One can use it to put data into objects; for example, `mynames <- c(''workshop,'' ''gender,'' ''q1,'' ''q2,'' ''q3,'' ''q4'')`. The combine function, `c`, puts the names together into a single object called a character vector. R also uses the double equal signs, `==`, as a logical comparison operator, as in, `gender == ''f.''` The # operator is used to begin a comment, which then continues until the end of the line. Semicolons can be used to enter multiple commands on a single line, as in SAS, but they are not usually used to end a command.

Commands can begin and end anywhere on a line, and additional spaces are ignored. The user can continue a command on a new line so long as the fragment left behind is not already a complete command itself. Going to a new line after a comma is usually a safe dividing point, and R commands are filled with commas.

One of the simplest R functions is `mean`. Typing `help(mean)` provides more detail on the form of the function: `mean(x,trim=0,na.rm=FALSE)`, where `x` is a data frame, a numeric vector, or a date. *Trim* is the "argument" that tells R what percentage of the extreme values to exclude before calculating the mean. The zero indicates the default value; that is, none of the data will be trimmed unless the user changes that percentage. The `na.rm` argument appears in many R functions. R uses `NA` to represent "Not Available," or missing values, as mentioned earlier. The default value of "`FALSE`" indicates that R will not remove (`rm`) them unless the user tells it otherwise. One must be careful because, as noted earlier, any statistical calculations will result in `NA` if any missing values are present. Many functions in R are set this way by default, a feature that is the opposite of SAS and SPSS, which almost always assume the user wants to use all the data unless the user specifies otherwise.

If one of the variables in the data set is categorical or string, R will compute the result but provide a warning. If users list every argument in order, they need not name them all. However, most people skip naming the first argument and then name the others and include them only if they wish to change their default values. For example, `mean(mydata,na.rm=TRUE)`.

Unlike SAS or SPSS, the output in R does not appear nicely formatted and ready to publish. However the user can use the functions in the *prettyR* and *Hmisc* packages to make the results of tabular output more ready for publication.

### Strengths and Weaknesses

R is coherent and intuitive for many users. Once users learn how to use the program, they will find

it easy to program new statistical methods. The graphics of the language allow easy production of advanced, publication-quality graphics. Since a wide variety of experts use the program, R includes a comprehensive library of statistical functions, including many cutting-edge statistical methods. In addition to this, many third-party specialized methods are publicly available. And most important, R is free and open source.

A common concern of beginning users of R is the steep learning curve involved in using it. Such concern stems from the fact that R is a command-driven environment. Consequently, the statistical analysis is performed in a series of steps, in which commands are typed out and the results from each step are stored in objects that can be used by further inquiries. This is contrary to other programs, such as SPSS and SAS, which require users to determine all characteristics of the analysis up front and provide extensive output, thus relying on the users to identify what is relevant to their initial question.

Another source of complaints relates to the difficulty of writing new functions. The more complex the function, the more difficult it becomes to identify errors in syntax or logic. R will prompt the user with an error message, but no indication is given of the nature of the problem or its location within the new code. Consequently, despite the advantage afforded by being able to add new functions to R, many users may find it frustrating to write new routines. In addition, complex analyses and simulations in R tend to be very demanding on the computer memory and processor; thus, the more complex the analysis, the longer the time necessary to complete the task, sometimes days.

Large data sets or complex tasks place heavy demands on computer RAM, resulting in slow output.

*Brandon K. Vaughn and Aline Orr*

***See also*** SAS; SPSS; Statistica; SYSTAT

**Websites**

Comprehensive R Archive Network (CRAN): http://CRAN.R-project.org
The R Project for Statistical Computing: http://www.r-project.org

# $R^2$

$R$-squared ($R^2$) is a statistic that explains the amount of variance accounted for in the relationship between two (or more) variables. Sometime $R^2$ is called the coefficient of determination, and it is given as the square of a correlation coefficient.

Given paired variables $(X_i, Y_i)$, a linear model that explains the relationship between the variables is given by

$$Y = \beta_0 + \beta_1 X + e,$$

where $e$ is a mean zero error. The parameters of the linear model can be estimated using the least squares method and denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively. The parameters are estimated by minimizing the sum of squared residuals between variable $Y_i$ and the model $\beta_0 + \beta_1 X_i$, that is,

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0,\beta_1}{\operatorname{argmin}} (Y_i - \beta_0 + \beta_1 X_i)^2.$$

It can be shown that the least squares estimations are

$$\hat{\beta}_0 = \overline{Y} - \overline{X}\frac{S_{xy}}{S_{xx}} \quad \text{and} \quad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}},$$

where the sample cross-covariance $S_{xy}$ is defined as

$$S_{xy} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \overline{XY} - \overline{XY}.$$

Statistical packages such as SAS, SPLUS, and R provide a routine for obtaining the least squares estimation. The estimated model is denoted as

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X.$$

With the above notations, the sum of squared errors (*SSE*), or the sum of squared residuals, is given by

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

*SSE* measures the amount of variability in $Y$ that is not explained by the model. Then how does one measure the amount of variability in $Y$ that is

explained by the model? To answer this question, one needs to know the total variability present in the data. The total sum of squares (*SST*) is the measure of total variation in the *Y* variable and is defined as

$$SST = \sum_{i=1}^{n} (Y_i - \overline{Y})^2,$$

where $\overline{Y}$ is the sample mean of *Y* variables, that is,

$$\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i.$$

Since *SSE* is the minimum of the sum of squared residuals of any linear model, *SSE* is always smaller than *SST*. Then the amount of variability explained by the model is *SST* − *SSE*, which is denoted as the regression sum of squares (*SSR*), that is,

$$SST = SST - SSE.$$

The ratio *SSR*/*SST* = (*SST* − *SSE*)/*SST* measures the proportion of variability explained by the model. The coefficient of determination ($R^2$) is defined as the ratio

$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SSE}.$$

The coefficient of determination is given as the ratio of variations explained by the model to the total variations present in *Y*. Note that the coefficient of determination ranges between 0 and 1. $R^2$ value is interpreted as the proportion of variation in *Y* that is explained by the model. $R^2 = 1$ indicates that the model exactly explains the variability in *Y*, and hence the model must pass through every measurement $(X_i, Y_i)$. On the other hand, $R^2 = 0$ indicates that the model does not explain any variability in *Y*. $R^2$ value larger than .5 is usually considered a significant relationship.

## Case Study and Data

Consider the following paired measurements from Moore and McCabe (1989), based on occupational mortality records from 1970 to 1972 in England and Wales. The figures represent smoking rates and deaths from lung cancer for a number of occupational groups.

| *Smoking index* | *Lung cancer mortality index* |
|---|---|
| 77 | 84 |
| 137 | 116 |
| 117 | 123 |
| 94 | 128 |
| 116 | 155 |
| 102 | 101 |
| 111 | 118 |
| 93 | 113 |
| 88 | 104 |
| 102 | 88 |
| 91 | 104 |
| 104 | 129 |
| 107 | 86 |
| 112 | 96 |
| 113 | 144 |
| 110 | 139 |
| 125 | 113 |
| 133 | 146 |
| 115 | 128 |
| 105 | 115 |
| 87 | 79 |
| 91 | 85 |
| 100 | 120 |
| 76 | 60 |
| 66 | 51 |

For a set of occupational groups, the first variable is the smoking index (average 100), and the second variable is the lung cancer mortality index (average 100). Suppose we are interested in determining how much the lung cancer mortality index (*Y* variable) is influenced by the smoking index (*X* variable). Figure 1 shows the scatterplot of the smoking index versus the lung cancer mortality index. The straight line is the estimated linear model, and it is given by

$$Y = -2.8853 + 1.0875X.$$

*SSE* can be easily computed using the formula

$$SSE = \sum_{i=1}^{n} Y_i^2 - \hat{\beta}_0 \sum_{i=1}^{n} Y_i - \hat{\beta}_1 \sum_{i=1}^{n} X_i Y_i, \qquad (1)$$

and *SST* can be computed using the formula

$$SST = \sum_{i=1}^{n} Y_i^2 - \frac{1}{n} \left( \sum_{i=1}^{n} Y_i \right)^2. \qquad (2)$$

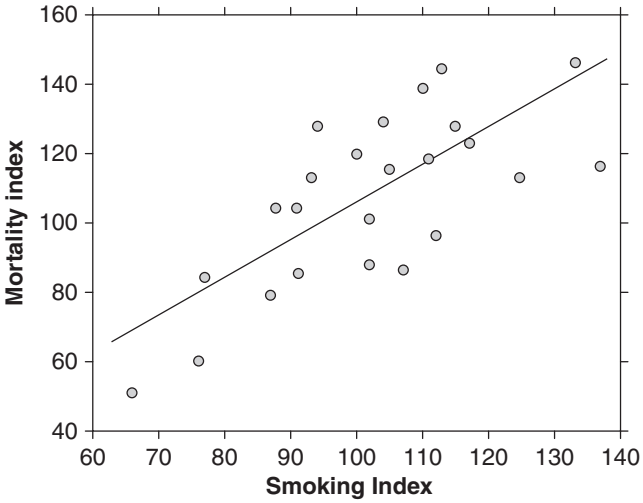



**Figure 1** Scatterplot of Smoking Index Versus Lung Cancer Mortality Index

*Note:* The straight line is the linear model fit obtained by the least squares estimation.

In this example the coefficient of determination is .5121, indicating that the smoking index can explain the lung cancer mortality index.

### Relation to Correlation Coefficient

With the previous Equations 1 and 2, $R^2$ can also be written as a function of the sample cross-covariance:

$$SSE = nS_{yy} - n\frac{{S_{xy}}^2}{S_{xx}} \quad \text{and} \quad SST = nS_{yy}.$$

Then the coefficient of determination can be written as

$$\mathrm{R}^2 = \frac{{S_{xy}}^2}{S_{xx}S_{yy}} = \frac{\left(\overline{XY} - \overline{X}\,\overline{Y}\right)^2}{(\overline{X^2} - \overline{X}^2)(\overline{Y^2} - \overline{Y}^2)},$$

which is the square of the Pearson product-moment correlation coefficient.

$$\mathrm{R} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})\sum_{i=1}^{n}(Y_i - \overline{Y})}}.$$

In the above example, the correlation coefficient is .7162, so the correlation square is .5121, the $R^2$ value.

### $R^2$ for General Cases

The definition of the coefficient of determination can be further expanded in the case of multiple regression. Consider the following multiple regression model:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + e,$$

where $Y$ is the response variable and $X_1, X_2, \cdots, X_p$ are $p$ regressors, and $e$ is a mean zero error. The unknown parameters $\beta_1, \beta_2, \cdots, \beta_p$ are estimated by the least squares method. The sum of squared residuals is given by

$$\begin{aligned} SSE &= \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \\ &= \sum_{i=1}^{n}\left[Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p)\right]^2, \end{aligned}$$

while the total sum of squares is given by

$$SST = \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2.$$

Then the coefficient of multiple determination is given by

$$R^2 = \frac{SST - SSE}{SST},$$

which is the square of the multiple correlation coefficient $R$. As the number of regressors increases, the $R^2$ value also increases, so $R^2$ cannot be a useful measure for the goodness of model fit. Therefore, $R^2$ is adjusted for the number of explanatory variables in the model. The adjusted $R^2$ is defined as

$$R^2_{\text{adj}} = 1 - (1 - R^2)\frac{n-1}{n-p-1} = \frac{(n-1)R^2 - p}{n-p+1}.$$

It can be shown that $R^2_{\text{adj}} \le R^2$. The coefficient of determination can be further generalized in more general cases using the likelihood method.

*Moo K. Chung*

*See also* Coefficients of Correlation, Alienation, and Determination; Correlation; Pearson Product-Moment Correlation Coefficient

**Further Readings**

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population. *Biometrika, 10,* 507–521.
Moore, D. S., & McCabe, G. P. (1989). *Introduction to the practice of statistics.* New York: W. H. Freeman.
Nagelkerke, N. J. D. (1992). Maximum likelihood estimation of functional relationships. *Lecture Notes in Statistics, 69,* 110.

# Radial Plot

The radial plot is a graphical method for displaying and comparing observations that have differing precisions. Standardized observations are plotted against the precisions, where precision is defined as the reciprocal of the standard error. The original observations are given by slopes of lines through the origin. A scale of slopes is sometimes drawn explicitly.

Suppose, for example, that data are available on the degree classes obtained by students graduating from a university and that we wish to compare, for different major subjects, the proportions of students who achieved upper second-class honors or higher. Typically, different numbers of students graduate in different subjects. A radial plot will display the data as proportions so that they may be compared easily, allowing for the different sample sizes. Similarly, a radial plot can be used to compare other summary statistics (such as means, regression coefficients, odds ratios) observed for different sized groups, or event rates observed for differing time periods.

Sometimes, particularly in the natural and physical sciences, measurements intrinsically have differing precisions because of natural variation in the source material and experimental procedure. For example, archaeological and geochronological dating methods usually produce an age estimate and its standard error for each of several crystal grains or rock samples, and the standard errors differ substantially. In this case, the age estimates may be displayed and compared using a radial plot in order to examine whether they agree or how they differ. A third type of application is in *meta-analysis,* such as in medicine, to compare estimated treatment effects from different studies. Here the precisions of the estimates can vary greatly because of the differing study sizes and designs. In this context the graph is often called a *Galbraith plot.* In general, a radial plot is applicable when one wants to compare a number of estimates of some parameter of interest, for which the estimates have different standard errors.

A basic question is, Do the estimates agree (within statistical variation) with a common value? If so, what value? A radial plot provides a visual assessment of the answer. Also, like many graphs, it allows other features of the data to be seen, such as whether the estimates differ systematically in some way, perhaps due to an underlying factor or mixture of populations, or whether there are anomalous values that need explanation. It is inherently not straightforward to compare individual estimates, either numerically or graphically, when their precisions vary. In particular, simply plotting estimates with error bars does not allow such questions to be assessed.

The term *radial plot* is also used for a display of *directional data,* such as wind directions and velocities or quantities observed at different times of day, via radial lines of different lengths emanating from a central point. This type of display is not discussed in this entry.

## *Mathematical Properties*

Let $z_1, z_2, \ldots, z_n$ denote $n$ observations or estimates having standard errors $\sigma_1, \sigma_2, \ldots, \sigma_n$, which are either known or well estimated. Then we plot the points $(x_i, y_i)$ given by $x_i = 1/\sigma_i$ and $y_i = (z_i - z_0)/\sigma_i$, where $z_0$ is a convenient reference value. Each $y_i$ has unit standard deviation, so each point has the same standard error with respect to the $y$ scale, but estimates with higher precision plot farther from the origin on the $x$ scale. The (centered) observation $(z_i - z_0)$ is equal to $y_i/x_i$, which is the slope of the line joining (0, 0) and $(x_i, y_i)$, so that values of $z$ can be shown on a scale of slopes. Figure 1 illustrates these principles.

Furthermore, if each $z_i$ is an unbiased estimate of the *same* quantity $\mu$, say, then the points will scatter with unit standard deviation about a line

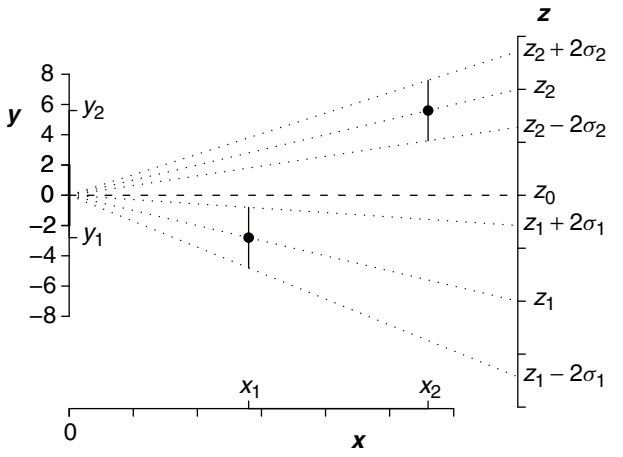



**Figure 1** The Geometry of a Radial Plot Showing Linear Scales for *x*, *y*, and *z* and Two Points With Different Precisions

*Note:* All points have unit standard deviation on the *y* scale, but higher precision points generate shorter confidence intervals on the *z* scale.

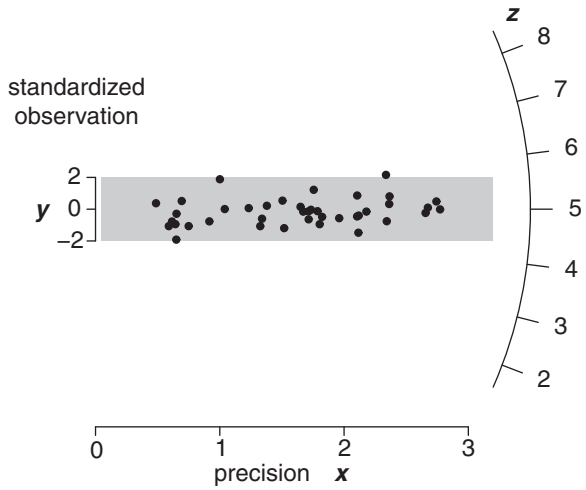



**Figure 2** A Radial Plot of $n = 40$ Random Values $z_1, z_2, \ldots, z_n$, Where zi Is From a Normal Distribution With Mean $\mu = 5$ and Standard Deviation $\sigma_i$, Where $1/\sigma_i$ Is Randomly Sampled Between 0.4 and 2.8

*Note:* The points scatter homoscedastically, nearly all within $\pm 2$ of the horizontal radius corresponding to $z_0 = \mu = 5$.

from (0, 0) with slope $\mu - z_0$. In particular, points scattering with unit standard deviation about the *horizontal* radius agree with the reference value $z_0$. This provides a simple visual assessment of how well several estimates agree with each other or with a reference value: points should scatter roughly equally about a radius and nearly all within a $\pm 2$ band. This is illustrated in Figure 2 using some simulated data with $\mu = z_0 = 5$. Such assessment can be applied to all of the points or to specific subsets of them. To facilitate this, it is useful to highlight the range $\pm 2$ on the *y* axis, or even to show just this range, so it acts as a "two-sigma" error bar applicable to any point.

For the specific purpose of assessing homogeneity, it is not necessary to draw the *z* scale, but doing so allows one to interpret the points in context. Also, it is easier to visualize radii if the scale is drawn as an arc of a circle, as in Figure 2. This is no longer a linear scale (i.e., equal divisions of *z* are not quite equal divisions of arc length), but this is of no disadvantage because we want to make position judgments on the *z* scale rather than difference judgments (the latter being complicated due to the differing precisions). In general, $z_0$ should be chosen so that the horizontal radius goes through the middle of the data—a common choice being the weighted average $z_0 = \sum w_i z_i / \sum w_i$ where $w_i = 1/\sigma_i^2$—though sometimes other choices of $z_0$ may be appropriate. When each $z_i$ has the same expectation $\mu_i$, this weighted average is the optimal (minimum variance unbiased) estimate of $\mu$.

There is a close connection with fitting a regression line through the origin to the $(x_i, y_i)$ data. The ordinary (unweighted) least squares estimate of the slope is $\sum x_i y_i / \sum x_i^2$, which is the same as $\sum w_i(z_i - z_0)\sum w_i$. Furthermore, when $z_0 = \sum w_i z_i / \sum w_i$, this slope is zero, so all fitted values are zero, and plotting $y_i$ against $x_i$ is the same as plotting the residuals against $x_i$.

### *Transformations*

Often it is appropriate to plot estimates (by whatever method) with respect to a transformed scale, particularly if on the original scale the precision is related to the size of the estimate. Sometimes the standard error of an estimate is approximately proportional to the size of the estimate (or of the quantity being estimated). Then one would use a log transformation—that is, take $z_i$ to be the natural log of the *i*th estimate

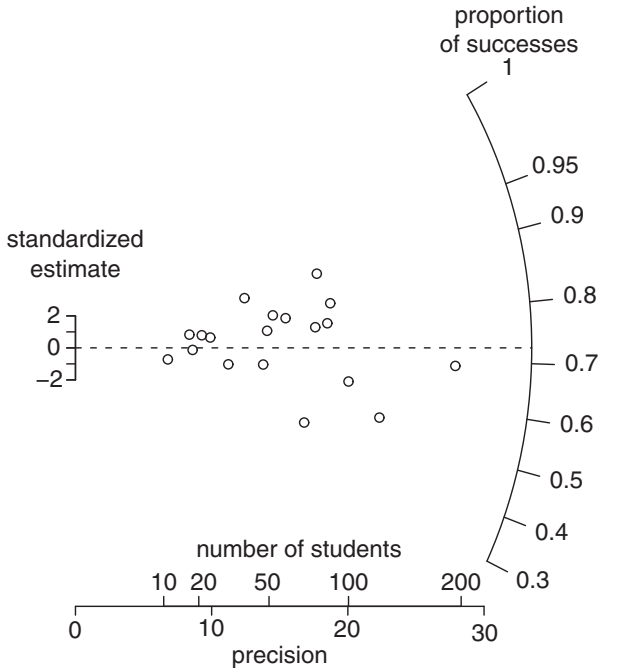



**Figure 3** A Radial Plot of Proportions of Students Graduating With Upper-Second-Class Honors or Better for 19 Degree Subjects

*Note:* The horizontal radius corresponds to the overall proportion for all students.

and $\sigma_i$ to be the standard error of $z_i$, so that $z_i$ and $\sigma_i$ are approximately independent. In this case, $\sigma_i$ approximately equals the *relative* standard error of the original estimate $e^{z_i}$. Nonlinear scales can be used to show the original measurements and their relative standard errors.

To compare *odds ratios* of binary event outcomes (e.g., for a given treatment compared to a control, for each of several studies) one would normally use a log scale (i.e., let $z_i$ be the log odds ratio for the $i$th study). For the radial plot, one would choose $z_0 = 0$, so the horizontal radius indicates an odds ratio of 1, corresponding to equal risk for treatment and control, and a positive (negative) slope indicates a higher (lower) event probability for the treatment group. The same applies to *risk ratios* (or relative risks).

Other useful transformations include the *square root,* for comparing event rates estimated from Poisson counts, and the *angular* transformation for comparing binomial proportions. In the former case, suppose $r_i$ events are counted in time $t_i$ where $r_i$ is from a Poisson distribution with mean $\lambda_i t_i$. Then the estimate $r_i/t_i$ of $\lambda_i$ has standard error $\sqrt{\lambda_i/t_i}$, but the transformed estimate, $z_i = \sqrt{r_i/t_i}$, has approximate standard error $\sigma_i = 1/\sqrt{4t_i}$, which does not depend on $\lambda_i$.

Likewise, suppose $r_i$ is from a binomial distribution with index $n_i$ and success probability $\theta_i$. Then the standard error of $r_i/n_i$ is $\sqrt{\theta_i(1-\theta_i)/n_i}$, which depends on $\theta_i$, but the transformed estimate

$$z_i = \arcsin\sqrt{\frac{r_i + \frac{3}{8}}{n_i + \frac{3}{4}}}$$

has approximate standard error $\sigma_i = 1/\sqrt{4n}$, which does not depend on $\theta_i$. The correction terms 3/8 and 3/4 make practically no difference for moderate or large counts but allow one to plot estimates sensibly when $r_i = 0$ or $n_i$. Of course, using a point estimate and standard error in such cases (rather than a confidence interval) is necessarily rough. With the advent of generalized linear models, such transformations are not used much nowadays for statistical modeling, but they are still useful for graphical analysis.

### *Examples*

Figure 3 illustrates a radial plot of proportions using the angular transformation. The data are, for each of 19 degree subjects, the number of students $n_i$ graduating, along with the number $r_i$ of these who achieved upper-second-class honors or better. The students all graduated from a particular UK university in a 6-year period, all were under 21 years of age on entry, and all had achieved the same entry qualification of 26 A-level points.

In Figure 3, the slope axis shows a nonlinear scale of proportions (rather than $z$), and the $x$ axis shows numbers of students (in addition to $1/\sigma$). The overall proportion of successes is 0.73, and $z_0$ is chosen to equal $\arcsin\sqrt{0.73}$. The graph shows clear heterogeneity (with respect to binomial variation) in the success rates for different subjects—even though the students all had the same A-level score—with several points well outside the $\pm 2$ horizontal band. This figure could be used as a basis for further discussion or analysis.

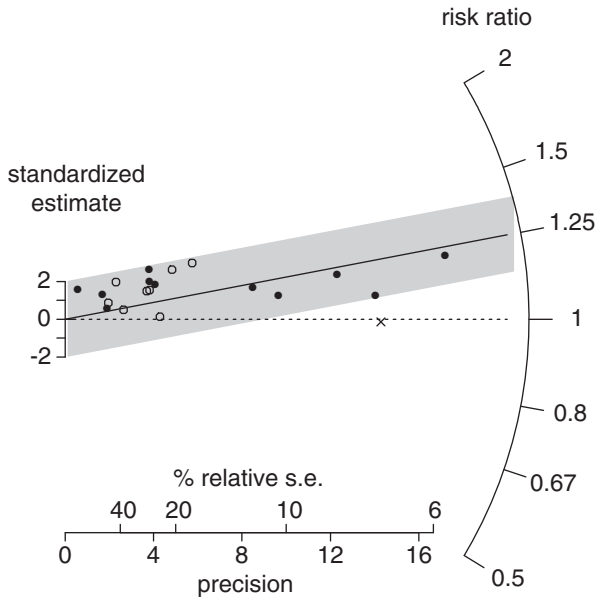



**Figure 4** A Radial Plot of Estimated Risk Ratios From a Meta-Analysis of the Effect of Passive Smoking on Coronary Heart Disease

*Source:* Data from J. He et al., 1999, Passive smoking and the risk of coronary heart disease: A meta-analysis of epidemiologic studies. *New England Journal of Medicine,* 340, 920–926.

*Notes:* Filled circles denote cohort studies, and open circles denote case–control studies. The cross is from a rogue study omitted from the meta-analysis.

Figure 4 displays a summary of results from a meta-analysis of studies of the possible effect of passive smoking on coronary heart disease (CHD). There are 10 cohort studies, identified by filled circles, and 8 case–control studies, identified by open circles. The cross indicates the result from an 11th cohort study that was omitted from the original analysis. The outcome variable was death due to CHD (within a given time), and the risk ratio (or relative risk) comparing individuals exposed to passive smoking with those not so exposed was estimated for each study, adjusting for age, sex, and some medical factors that were not the same in all studies. The basic data for plotting are ($z_i$, $\sigma_i$) the estimate and standard error of the natural logarithm of the risk ratio.

Figure 4 shows the actual risk ratios on a logarithmic scale, centered at $z_0 = 0$ corresponding to a risk ratio of 1. The $x$ scale shows, in addition to the precisions, some values of $100\sigma$, which here indicate percentage relative standard errors of the risk ratios. These features are included for illustration and are of course not essential to the basic plot.

The graph reveals some interesting patterns. Except for the cross (the omitted study), all of the estimated risk ratios are greater than 1, suggesting that exposure to passive smoking is associated with a higher risk of death from CHD. Five of the points are from large cohort studies and have fairly precise estimates, but the estimates for the remaining studies are rather imprecise. Some have very high estimated risk ratios (off the scale), with relative standard errors much greater than 20%. Nevertheless, *all* these estimates lie within the $\pm 2$ shaded band, apparently agreeing with a common risk ratio of about 1.25. However, the filled points by themselves (i.e., results for the cohort studies) do not scatter randomly about this radius—those for the small studies are all above the line (corresponding to a higher risk ratio), and those for the larger studies are below it (corresponding to a lower risk ratio than 1.25). It has been suggested that this may be due to *publication bias,* resulting in a lack of small studies with risk ratios close to 1.

*Rex Galbraith*

***See also*** Scatterplot; Standard Error of Estimate

**Further Readings**

Armitage, P., Berry, G., & Matthews, J. N. S. (2002). *Statistical methods in medical research* (4th ed.). Oxford, UK: Blackwell.

Galbraith, R. F. (1988). Graphical display of estimates having differing standard errors. *Technometrics, 30,* 271–281.

Galbraith, R. F. (1988). A note on graphical presentation of estimated odds ratios from several clinical trials. *Statistics in Medicine, 7,* 889–894.

Galbraith, R. F. (1994). Some applications of radial plots. *Journal of the American Statistical Association, 89,* 1232–1242.

# Random Assignment

Random assignment is the process by which researchers select individuals from their total

sample to participate in a specific condition or group, such that each participant has a specifiable probability of being assigned to each of the groups or conditions. These different conditions or groups represent different levels of the independent variable. Random assignment is generally considered the most important criterion to qualify a research design as an *experimental* design. A common example of random assignment is a medical trial in which a researcher assigns some participants at random to receive either a treatment drug or a placebo (a pill that looks like the medication but is known to be inert). Because of the random assignment, any differences that are observed in the outcome variable can, under certain assumptions, be causally attributed to the independent variable. Random assignment is not to be confused with *random selection,* which is the process by which researchers select a smaller sample from the larger population.

This entry discusses issues of causality, specifically how random assignment is used to help ensure that the observed differences between groups are due to the manipulated independent variable and not other preexisting differences between the groups. It also provides examples of different methods of randomly assigning participants to groups and describes the different levels of randomization, such as random assignment at the school level as opposed to the individual level. Finally, this entry discusses some potential problems that may arise with random assignment, particularly when one is working with human participants.

## Causality and Internal Validity

The most important tenet of establishing causality between two variables is that there must be no other plausible explanation for the observed relationship between the variables. That is, in order to validly make the claim that independent variable A causes changes in outcome variable B, all other potential causes of changes in B must be ruled out. The many other potential variables that may affect the outcome variable are referred to as *confounding variables* or *nuisance variables*. If one is able to effectively rule out every other explanation for the relationship between variables A and B, the study has good *internal validity.* Randomly assigning participants to the various levels of the independent variable increases internal validity, and thus aids in establishing causation, because it helps create the "everything else equal" condition: The randomization process roughly equates the groups on *every* potential confounding variable. Thus, any differences observed between the groups over and above the small amount expected after random assignment must be due to the independent variable. The beauty of random assignment is that, if sample size is sufficiently large, it assures that all determinants of B, even unknown and unspecified ones, are largely evenly distributed between the groups.

Consider the following example of a researcher who is interested in the effects of an herbal medication on depression. From a pool of 100 depressed patients, the researcher randomly assigns half to receive the medication and the other half to receive a placebo pill. After 6 weeks, the researcher compares the levels of depression in the two groups. Because the researcher has randomly assigned participants to the two groups, any preexisting differences in the levels of depression, or any other possible confounding variable, should be approximately equated in these two groups. So if the difference in the levels of depression following treatment exceeds the small amount to be expected after randomization, the researcher can validly claim the difference is due to the medication.

An alternative to random assignment that is often used to control or equate for preexisting differences between groups is *matching.* Generally, though, when it is possible, the use of random assignment is much preferable to matching because it is impossible to match participants on every possible confounding variable.

The ability of random assignment to allow unambiguous attribution of causality to the independent variable can still be compromised by various imperfections in the way the study is carried out. Whenever the groups are treated differently in respects *other than* the independent variable, confounds can still exist. One well-known such problem involves *experimenter effects,* which can arise because the researcher knows which individuals are in which groups and may intentionally or unintentionally treat them differently, which could be the true cause of any differences later found. In order to avoid any such bias in the results,

researchers often conduct *double-blind studies*. In these types of studies, the person recording the results (as well as the participants themselves) does not know to which type of treatment each person is assigned and may also be naive as to the hypotheses being tested. This measure helps ensure equal treatment of the participants in the different conditions, as well as equal recording of results.

## How to Randomly Assign Participants

There are many methods that may be employed to randomly assign participants. These methods range from the simple, such as flipping a coin, to the more complex, such as using random number generators. It does not generally matter which method of random assignment is used, as long as each participant in fact has the prespecified probability of being assigned to each of the experimental conditions (which is typically, but by no means necessarily, an *equal chance* of being assigned to each condition). If a researcher needs to assign participants to one of two possible conditions, the researcher might flip a coin. If there are six conditions, the researcher might roll a six-sided die. It is common for researchers to use samples of college undergraduates as participants. In this case, the researcher could use the digits in participants' student identification numbers to randomly assign them to groups. More complex methods of random assignment include use of random number generators (for example, the RAND() function in Excel) or random number tables to assign participants to groups.

## Levels of Randomization

Depending on the type of research question in which one is interested, random assignment may occur at a variety of different levels. This entry has thus far provided examples of random assignment at the individual level, as when a researcher randomly assigns participants to receive either a medication or a placebo. It may also be of interest or necessity to assign participants at the group level rather than the individual level. For example, suppose a researcher is interested in studying the effects of an intervention to prevent tobacco use among high school students. Rather than randomly assigning individual high school students to receive either the intervention or the control, the researcher could randomly assign whole classrooms or even entire high schools to either the intervention program or the control. This level of random assignment may be preferred for some research because it reduces the likelihood of students' sharing information about the intervention to those assigned to the control condition. Further, randomly assigning predefined groups, such as schools, can often be more cost-effective than randomly assigning at the individual level. Randomly assigning at the group level does, however, introduce statistical issues, such as clustering effects, which need to be addressed through appropriate data analysis methods, such as multilevel modeling.

## Potential Randomization Problems

### *Unhappy Randomization*

Although the process of randomly assigning participants to groups should approximately equate the groups on all potential confounding variables, this will not occur in a small proportion of studies, and is more of a problem when sample sizes are small. Some inequality will always remain, but statistical tests take this error into account. In the small proportion of studies in which, despite proper randomization procedures, the groups are found to be unequal beyond this tolerable degree on one or more confounding variables, the condition is called *unhappy randomization.*

### *Compromised Randomization*

Program staff can compromise randomization by not strictly following its assignment rules. For example, in the Canadian National Breast Screening Study, it was feared (but fortunately not found) that nurses may have arranged for particular women to be enrolled in a condition other than the one designated by the randomization procedure. Experienced evaluators recognize the danger that program staff members may well prefer to use their own strategies for assigning conditions (for example, first-come–first-served or assignment based on the participants' appearance of need or motivation) rather than assignment at random. Thus the random assignment process needs to be

implemented in a way that prevents intake staff from manipulating it.

### *Treatment Noncompliance*

Issues arise when participants do not participate in the group to which they were assigned. This occurs when, for a variety of reasons (for example, the experience of side effects, the amount of effort or diligence required for compliance), participants choose not to actually follow the regimen to which they are assigned. This noncompliance is especially a problem when the conditions are unequal in their propensity to elicit noncompliance. It becomes a threat to internal validity because these participants are essentially self-selecting out of their group and possibly into another. There are statistical procedures that address this issue, such as *intent-to-treat analysis* and newer methods such as *propensity score analysis*.

### *Randomizing the Invitation, Not the Treatment*

A related problem occurs when the researcher randomly decides which condition to *invite* participants to (rather than the one they will receive), and those who decline are simply left out of the analysis. When the considerations leading to the decision to participate are different between the conditions, serious bias can enter. Consider a study in which ill individuals are to receive either an inert placebo or a drug with formidable side effects, and only those who are randomized to the drug condition are informed about the side effects. Assume a high proportion of these (but not the placebo control) thus decline to participate. Such an occurrence would alter the composition of the two conditions and render them not comparable. The solution to preserve pretreatment comparability of the groups and therefore internal validity is to explain the requirements of both conditions, including the side effects, to everyone and obtain participation only from those who can commit to either condition, whichever arises randomly.

While internal validity is preserved in such a procedure, large numbers of ill people may exclude themselves from the study. Thus, the study may have a sample that underrepresents individuals with the illness. This limits the group about whom the conclusions are applicable. This limit on the ability to generalize the results is known as *external validity*. It has been well recognized that rigorous attempts to secure high internal validity often threaten external validity, and thus recognition of the trade-offs between the two is imperative.

### *Differential Attrition*

There are other situations in which participants may just drop out of the study altogether. This is known as attrition, and it can be a threat to internal validity, particularly if the reason the participants drop out is the condition to which they were assigned, therefore leading to differential attrition. Depending on the type and extent of the attrition, researchers may use a variety of statistical methods to deal with the missing data and salvage causal interpretation of results.

### *Ethical Considerations*

There are many instances in which random assignment cannot be employed for ethical reasons. For example, researchers may want to investigate whether smoking cigarettes causes lung cancer. Because there are known adverse health effects of cigarette smoking, as well as high risk of addiction associated with tobacco use, it would be unethical to randomly assign some human participants to smoke cigarettes and others to abstain. When it becomes unethical to randomly assign human participants to groups, researchers often use animals as subjects. Researchers might also resort to quasi-experimental designs, such as observing over time a group of people who choose to smoke as well as a comparison group of people who do not smoke. Although inferences of causation are less confidently made than with randomized designs, quasi-experiments that are carefully designed and analyzed may still yield results bearing on causal processes.

*Sanford L. Braver, Stephanie E. Moser, and Felix Thoemmes*

***See also*** Confounding; Experimental Design; Experimenter Expectancy Effect; External Validity; Internal Validity; Matching; Quasi-Experimental Design; Random Selection

**Further Readings**

Braver, S. L., & Smith, M. C. (1996). Maximizing both external and internal validity in longitudinal true experiments with voluntary treatments: The "combined modified" design. *Evaluation and Program Planning, 19*(4), 287–300.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Boston: Houghton Mifflin.

Gueron, J. M. (2002). The politics of random assignment: Implementing studies and affecting policy. In F. Mosteller & R. F. Boruch (Eds.), *Evidence matters: Randomized trials in education research*. Washington, DC: Brookings Institution Press.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. New York: Houghton Mifflin.

## Random-Effects Models

Random-effects models are statistical models in which some of the parameters (effects) that define systematic components of the model exhibit some form of random variation. Statistical models always describe variation in observed variables in terms of systematic and unsystematic components. In *fixed-effects models,* the systematic effects are considered fixed or nonrandom. In random-effects models, some of these systematic effects are considered random. Models that include both fixed and random effects may be called *mixed-effects models* or just *mixed models.*

Randomness in statistical models usually arises as a result of random sampling of units in data collection. When effects can have different values for each unit that is sampled, it is natural to think of them as random effects. For example, consider observations on a variable $Y$ that arise from a simple random sample from a population. We might write the model

$$Y_i = \mu + \varepsilon_i,$$

where $\mu$ is the population mean and $\varepsilon_i$ is a residual term to decompose the random observation into a systematic part (a fixed effect) $\mu$, which is the same for all samples, and an unsystematic part $\varepsilon_i$, which in principle is different for every observation that could have been sampled. Because the $i$th individual in the sample is a random draw from the population, the value $Y_i$ associated with that individual is random and so is $\varepsilon_i$. Thus in this simple model, $\mu$ is a fixed effect and $\varepsilon_i$ is a random effect.

More complex sampling designs typically lead to more complex statistical models. For example, two-stage (or multistage) sampling designs are quite common. In such designs, the sample is obtained by first sampling intact aggregate units such as schools or communities (generically called clusters in the sampling literature), then sampling individuals within those aggregate units. If our observations on $Y$ arise from a two-stage sample, then a natural model might decompose the variation in $Y$ into systematic parameters associated with the overall population mean ($\mu$), and the effect of the aggregate unit ($\xi$), in addition to the residual ($\varepsilon$). Because there are now two aspects of the data to keep track of (aggregate units and individuals within those units), we might use two subscripts, denoting the $j$th observation in the $i$th aggregate unit (the $i$th cluster) by $Y_{ij}$ and write the model as

$$Y_{ij} = \mu + \xi_i + \varepsilon_{ij}.$$

Here $\xi_i$, the effect of the $i$th aggregate unit, is a random effect because the aggregate units were chosen at random as the first stage of the sample. Similarly, the $\varepsilon_{ij}$s are also random effects because the individuals within the aggregate units were chosen at random in the second stage of the sample.

In most research designs, the values of the outcome variable associated with particular individuals (or residuals associated with particular individuals) are of little interest in themselves, because the individuals are sampled only as a means of estimating parameters describing the population. Similarly, in research designs that use multistage sampling, the specific values of the cluster-level random effects are of little interest (because inferences of interest are about the larger population). However, in both cases the variances associated with individual-level or cluster-level random effects are of interest. These variances, often called *variance components,* have an impact on significance tests and precision of estimates of fixed effects, and they play a key role in

determining the sensitivity and statistical power of research designs involving random effects. Consequently, random-effects models in some areas (such as analysis of variance) are often called *variance component models.*

To illustrate the role of the variance components on precision, consider our example of sampling from a population with mean $\mu$ and variance $\sigma_T^2$, which can be decomposed into a between-cluster variance $\sigma_B^2$ and a within-cluster variance $\sigma_W^2$, so that $\sigma_T^2 = \sigma_B^2 + \sigma_W^2$. If we estimate the population mean $\mu$ from the mean of a simple random sample of $N = mn$ individuals, the variance of that estimate (the sample mean) would be $\sigma_T^2/N$. If we estimate the population mean $\mu$ using a sample of the same size $N = mn$, but using a two-stage cluster sample of $m$ clusters of $n$ individuals each, the variance would be

$$\left(\sigma_T^2/N\right)\left[1 + (n-1)\left(\sigma_B^2/\sigma_T^2\right)\right],$$

so that the variance of the estimate is larger by a fraction (the quantity in square brackets) that depends on the between-cluster variance component.

This discussion refers to formal sampling designs (such as simple random sampling and two-stage sampling) for precision, but the principle that effects that are specific to sampled units (and may be different for each unit that might have been sampled) must be random effects applies whether or not formal sampling designs are used.

Several common random effects models are illustrated below, with an emphasis on the models, as opposed to data analysis procedures.

## Analysis of Variance Models for Experimental Designs

For many researchers, the most familiar examples of random-effects models are analysis of variance models that arise in conjunction with experimental designs involving random effects. This is illustrated with three simple designs that are frequently used in themselves and are often the foundation for more complex designs: the completely randomized design, the hierarchical design, and the randomized blocks design.

### *Completely Randomized Design*

In the completely randomized design, individuals are randomly assigned to treatments. The inferential question is whether all the treatments in a population of treatments have the same effect. For example, consider a population of stimuli generated by a computer according to a particular algorithm. One might want to test whether the average response to all the stimuli is the same. An experiment might select the treatments to be tested by taking a random sample of $A$ treatments from a population of possible treatments, then assigning $n$ individuals at random to each of the $A$ treatments and observing the outcome. The model for the outcome of the $j$th individual assigned to the $i$th treatment might be

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij},$$

where $\mu$ is the grand mean, $\alpha_i$ is the effect of the $i$th treatment, and $\varepsilon_{ij}$ is a residual for the $j$th individual assigned to the $i$th treatment. Because we think of the treatments as being a random sample from a population of treatments, we would think of the $\alpha_i$s (the treatment effects) as being a random sample from a population of treatment effects; in other words, the $\alpha_i$ are random effects in this model.

The particular treatments that happen to be in the sample to be evaluated in this experiment are of little interest in themselves. The object of inference is the population of treatments. Consequently, the hypothesis we wish to test is not about the particular $\alpha_i$ associated with this sample of $A$ treatments, but about the population from which they are sampled. That is, we wish to know whether all the $\alpha_i$s in the population (not just the experimental sample) of treatment effects are identical. That is, the variance of the population of $\alpha_i$s, $\sigma_\alpha^2$, is zero. Thus the null hypothesis in this random effects model is

$$H_0 : \sigma_\alpha^2 = 0.$$

This illustrates how variance components are used in formulating hypotheses in random-effects models.

### *Hierarchical Design*

In a hierarchical design, individuals belong to preexisting aggregate units (*blocks* in experimental design terminology or *clusters* in sampling terminology). In the hierarchical design, entire aggregate units (blocks) are assigned to the same treatment. Because entire blocks or clusters are randomized to treatments, this design is sometimes called a *cluster-randomized* or *group-randomized design*. In experimental design terminology, the blocks are *nested* within treatments. In this design, blocks are taken to be random, that is, sampled from a population of blocks. Treatments may be fixed (as in the case of a small set of treatments of interest) or random (as in the case of the completely randomized design discussed above).

For example, consider an experiment designed to compare two mathematics curricula. In this case, the treatment (curriculum) is a fixed effect. However suppose that the experiment involved sampling $2m$ schools with $n$ students in each school and assigned $m$ schools to use each curriculum. The model for $Y_{ijk}$, the outcome of the $k$th student in the $j$th school receiving the $i$th treatment, might be

$$Y_{ijk} = \mu + \alpha_i + \beta_{j:i} + \varepsilon_{ijk},$$

where $\mu$ is the grand mean, $\alpha_i$ is the effect of the $i$th treatment, $\beta_{j:i}$ is the effect of the $j$th school nested within the $i$th treatment, and $\varepsilon_{ijk}$ is a residual for the $k$th individual in the $j$th school assigned to the $i$th treatment. Because we think of the schools as being a random sample from a population of schools, we would think of the $\beta_{j:i}$ (the school effects) as being a random sample from a population of school effects; in other words, the $\beta_{j:i}$s are random effects in this model.

### *Randomized Blocks Design*

In the randomized blocks design, as in the hierarchical design, individuals belong to preexisting aggregate units (blocks or clusters). In the randomized blocks design, individuals within aggregate units (blocks) are assigned to the different treatments so that every treatment occurs within each block. Because the individuals within each block are matched by virtue of being in the same block, this design is sometimes called a *matched design*. In experimental design terminology, the blocks are *crossed* with treatments. In this design, blocks are usually taken to be random, that is, sampled from a population of blocks, but in some circumstances blocks may be taken to be fixed (when the blocks included in the experiment constitute the entire population of blocks that are of interest). Treatments may be fixed (as in the case of a small set of treatments of interest) or random (as in the case of the completely randomized design discussed above).

Consider the example of an experiment designed to compare two mathematics curricula. In this case, the treatment (curriculum) is a fixed effect. However, suppose that the experiment involved sampling $m$ schools with $2n$ students in each school and assigning $n$ students within each school to use each curriculum. The model for $Y_{ijk}$, the outcome of the $k$th student in the $j$th school receiving the $i$th treatment, might be

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk},$$

where $\mu$ is the grand mean, $\alpha_i$ is the effect of the $i$th treatment, $\beta_j$ is the effect of the $j$th school, $\alpha\beta_{ij}$ is the interaction of the $i$th treatment with the $j$th school, and $\varepsilon_{ijk}$ is a residual for the $k$th individual in the $j$th school assigned to the $i$th treatment. Because we think of the schools as being a random sample from a population of schools, we would think of the $\beta_j$ (the school effects) as being a random sample from a population of school effects; in other words, the $\beta_j$s are random effects in this model. Similarly, we would think of the $\alpha\beta_{ij}$ (the Treatment $\times$ School interaction effects) as being a random sample from a population of Treatment $\times$ School interaction effects; in other words, the $\alpha\beta_{ij}$s are random effects in this model.

## Hierarchical Linear Models

Hierarchical linear models (also called *multilevel models*) can be used to analyze a very wide range of data, including survey data that arise from multistage samples, and longitudinal data, and even as an alternative to analysis of variance for experiments with hierarchical or randomized blocks designs. Hierarchical linear models specify the overall model via separate models for the data at each stage of sampling. In principle, these models can have any number of levels, although two- and

three-level models are the most common. Two-level models are used to illustrate hierarchical linear models here.

### *Hierarchical Linear Models for Survey Data*

Consider a survey with a two-stage sample, such as sampling of students by first obtaining a sample of schools, then obtaining a sample of students within each of the schools. The Level-1 (within-school) model specifies the model for individual observations within each school. The Level-2 (between-school) model specifies the model for variation of school-specific parameters across schools. Because the Level-2 model describes parameters that depend on the particular schools sampled and the sample is random, these parameters are random effects.

The Level-1 model for $Y_{ij}$, the outcome of the $j$th student in the $i$th school, might say that $Y_{ij}$ depends on a linear model including one or more individual-level covariates (here we use only one covariate, $X$):

$$Y_{ijk} = \beta_{0i} + \beta_{1i}X_{ij} + \varepsilon_{ij},$$

so that $\beta_{0i}$ is the intercept in school $i$ and $\beta_{i1}$ is the slope of $Y$ on $X$ in school $i$. Because $\beta_{0i}$ and $\beta_{1i}$ are both specific to school $i$, and the particular schools observed arise from sampling schools at random, it is natural to think of $\beta_{0i}$ and $\beta_{1i}$ as random effects. The student level residual, $\varepsilon_{ij}$ is also random because it depends of the particular student within a school, which is also sampled at random.

The Level-2 model includes two equations, one for $\beta_{0i}$ and the other for $\beta_{0i}$, which may include one or more school-level covariates (here we include one covariate, $W$):

$$\beta_{0i} = \gamma_{00} + \gamma_{01}W_i + \eta_{0i}$$

and

$$\beta_{1i} = \gamma_{10} + \gamma_{11}W_i + \eta_{1i},$$

where $\eta_{0i}$ and $\eta_{1i}$ are school-level residuals that describe how different a particular school's intercept (in the case of $\eta_{0i}$) or slope (in the case of $\eta_{1i}$) is from the value that would be predicted based on $W$. Note that the covariate $W$ has only a single subscript (here $i$, referring to the school) because its value is specific to the Level-2 unit. In the Level-2 model, the regression coefficients $\gamma_{00}$, $\gamma_{01}$, $\gamma_{10}$, and $\gamma_{11}$ are all fixed effects. Because $\beta_{0i}$ and $\beta_{1i}$ are random effects, the residuals $\eta_{0i}$ and $\eta_{1i}$ must be random effects.

In some circumstances, one or more of the Level-1 parameters (e.g., $\beta_{1i}$) might be assumed to be constant across all Level-2 units (all schools). This is functionally the same as declaring that the Level-2 residuals associated with that coefficient (e.g., the $\eta_{1i}$s) are identically zero or, equivalently, that the residual has zero variance.

### *Hierarchical Linear Models for Longitudinal Data*

Hierarchical models can also be applied to the analysis of longitudinal data. In this situation, the Level-2 units are individuals and the Level-1 units are observations within individuals. Let $Y_{ij}$ be the measurement at the $j$th time of the $i$th person. Then the Level-1 model describes the growth trajectory of the outcome over time for each person individually. For example, a quadratic growth trajectory might lead to a Level-1 (within-individual) model such as

$$Y_{ij} = \beta_{0i} + \beta_{1i}X_{ij} + \beta_{2i}X_{ij}^2 + \varepsilon_{ij},$$

where here $X_{ij}$ is the time of the $j$th measurement of the $i$th person. In this Level-1 model, the coefficients $\beta_{0i}$, $\beta_{1i}$, and $\beta_{2i}$ describe the growth trajectory for the $i$th person and, in principle, can take different values for each individual. Because individuals are sampled at random, $\beta_{0i}$, $\beta_{1i}$, and $\beta_{2i}$ are random effects.

The Level-2 model in longitudinal studies might reflect the impact of individual characteristics (such as family background, gender, or treatments received) on their growth trajectories. Because three coefficients are used to describe the individual growth trajectories at Level 1, there are three Level-2 equations. With one Level-2 covariate, the Level-2 (between-individual) model could be

$$\beta_{0i} = \gamma_{00} + \gamma_{01}W_i + \eta_{0i},$$

$$\beta_{1i} = \gamma_{10} + \gamma_{11}W_i + \eta_{1i},$$

and

$$\beta_{2i} = \gamma_{20} + \gamma_{21}W_i + \eta_{2i},$$

where $\eta_{0i}$, $\eta_{1i}$, and $\eta_{2i}$ are individual-level residuals that describe how different a particular individual's intercept (in the case of $\eta_{0i}$), linear trend (in the case of $\eta_{1i}$), or quadratic trend (in the case of $\eta_{2i}$) are from the value that would be predicted based on $W$. In the Level-2 model, $W$ and the regression coefficients $\gamma_{00}$, $\gamma_{01}$, $\gamma_{10}$, $\gamma_{11}$, $\gamma_{20}$, and $\gamma_{21}$ are all fixed effects. Because $\beta_{0i}$, $\beta_{1i}$, and $\beta_{2i}$ are random effects, the residuals $\eta_{0i}$, $\eta_{1i}$, and $\eta_{2i}$ must be random effects.

In some circumstances, one or more of the Level-1 parameters (e.g., $\beta_{2i}$) might be assumed to be constant across all Level-2 units (all individuals), or modeled as such if the data contain too little information to obtain precise estimates of all the fixed effects. This is functionally the same as declaring that the Level-2 residuals associated with that coefficient (e.g., the $\eta_{2i}$s) are identically zero or, equivalently, that the residual has zero variance.

### Random Effects Models for Meta-Analysis

Meta-analysis involves the combination of information across independent research studies by representing the results of each study via a quantitative index of effect size (such as a standardized mean difference, correlation coefficient, or odds ratio) and combining these estimates. Let $\theta_i$ be the effect size parameter associated with the $i$th study and let $T_i$ be a sample estimate of $\theta_i$, so that

$$T_i = \theta_i + \varepsilon_i,$$

where $\varepsilon_i = T_i - \theta_i$ is a residual reflecting error of estimation of $\theta_i$ by $T_i$. Fixed-effects models for meta-analysis treat the $\theta_i$ as fixed constants. In a formal sense, this is equivalent to limiting the statistical inference to studies that are identical to those observed, but with different samples of individuals drawn from the same populations.

In contrast, random-effects models treat the sample of studies observed as a random sample from a larger population of studies, so the effect-size parameters are therefore random effects. For example, a random effects model might posit that the $\theta_i$s are a sample from a population of effect size parameters with mean $\theta_\bullet$ and variance $\tau^2$, so that

$$T_i = \theta_\bullet + \xi_i + \varepsilon_i,$$

where $\xi_i = \theta_i - \theta_\bullet$ is a study-specific random effect with mean zero and variance $\tau^2$. The object of the statistical analysis is to estimate the mean (and possibly the variance) of the population of effect sizes, or in other words, to estimate the fixed effect $\theta_\bullet$ (and possibly $\tau^2$).

*Larry V. Hedges*

***See also*** Experimental Design; Fixed-Effects Models; Hierarchical Linear Modeling; Longitudinal Design; Meta-Analysis; Randomized Block Design; Random Sampling

#### Further Readings

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. New York: Academic Press.

Hedges, L. V., & Vevea, J. L. (1998). Fixed and random effects models in meta-analysis. *Psychological Methods, 3,* 486–504.

Kirk, R. (1995). *Experimental design*. Belmont, CA: Brooks Cole.

Kish, L. (1965). *Survey sampling*. New York: Wiley.

Raudenbush, S. W. (1993). Hierarchical linear models and experimental design. In L. K. Edwards (Ed.), *Applied analysis of variance in behavioral science* (pp. 459–496). New York: Marcel Dekker.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models*. Thousand Oaks, CA: Sage.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis*. Oxford, UK: Oxford University Press.

Snijders, T., & Bosker, R. (1999). *Multilevel analysis*. Thousand Oaks, CA: Sage.

## Random Error

Random error, as with most error in research design is not "error" in the sense of being an obvious mistake. Rather, it is variability that occurs in data due simply to natural inconsistency that exists in the world. Random error does not alter the measured values of a variable in a consistent or systematic way, but it does contribute to an imperfect measure of a variable.

### Effects on Data

Random error contributes variability in all data collection measures, such that according to

*classical test theory,* the observed value of a variable ($x_0$) is not actually the "true" value of the variable, but instead it is the true value ($x_t$) of the variable *plus* the value of the error ($e$). This idea can be written as $x_0 = x_t + e$. The error term may be made up of both random error ($e_r$) and *systematic error* ($e_s$), which is error that alters the values of the data in a consistent manner. Thus,

$$x_0 = x_t + e,$$
$$e = e_r + e_s,$$
$$x_0 = x_t + e_r + e_s.$$

Random error is equally likely to make an observed value of a variable higher or lower than its true value. Although the random error can lead to erroneous conclusions about the values of individual data points, the impact of random error on the value of any one observed variable is generally canceled out by other imperfect observed values of those data. Thus when a sample of data is collected, it is assumed that the random variability will not consistently contribute to the values of the data in one direction. Even though the random error contributes variability to observed data, the overall mean of all the observed data is assumed to be the same as the mean of the true values of the measured variable. The random error simply creates more "noise" in the data; even so, the noise is centered on the true value of the variable.

Because data measures are never perfect, hypothesis testing consists of conducting statistical analyses to determine whether the variability that exists in data is simply due to random error or whether it is due to some other factor, such as systematic error, or the effect of an independent variable. In hypothesis testing, the variability in the samples is compared with the estimated variability in the population in order to determine whether it is probable that the sample variability is equivalent to the variability expected due to random error. If there is more variability than one would expect simply due to random error, then the variance is said to be "statistically significant."

## Expected Value

When multiple samples are taken from a population, the values of the statistics calculated from those samples are assumed to contain random error. However, the mean of random error in those samples is assumed to be zero. In addition, because the random error is equally likely to increase or decrease the observed value of the true variable, and it is not expected to have a systematic impact, random error is generally assumed to have a normal distribution. If multiple samples are taken from a normal distribution with a mean of zero, then the mean of those samples will also be zero. Thus, given that the mean of the normally distributed random error distribution is zero, the expected value of random error is zero. Consequently, the mean value of the measured data is equal to the mean value of the true data plus the mean random error, which is zero. Thus even though random error contributes to data measures, the mean of the measured data is unaffected by the random error and can be expected to be equal to its true value.

## Causes and Related Terms

Random error can be caused by numerous things, such as inconsistencies or imprecision in equipment used to measure data, in experimenter measurements, in individual differences between participants who are being measured, or in experimental procedures. Random error is sometimes incorrectly referred to as experimental error or measurement error. *Experimental error* or *measurement error* can be due to both random error and systematic error. If an experiment is repeatedly performed, the effects of random error will differ with repetitions, while the effects of systematic error will likely be consistent across repetitions.

## Example

If researchers are conducting a study on the impact of caffeine on reaction times, they may collect reaction time data by having participants respond to a stimulus on a computer screen by pressing the space bar. Sometimes a participant may not be fully paying attention and will respond slightly slower than his or her normal reaction time. On other trials the participant may anticipate the stimulus and respond more quickly than the normal reaction time. These variations in response times are considered random error. Likewise if some of

the participants in the experiment were short on sleep the night before the experiment, their responses might be slower than if they were well rested. At the same time, the responses of well-rested participants might be faster than if they were not so well rested. These variations due to temporary differences in an individual's physical or emotional states are also considered random error. However, if all the participants in the experiment could not easily reach the space bar, causing all their responses to be recorded as slower than their initial attempts to respond, then that error in measurement would be systematic error.

*Jill H. Lohmeier*

***See also*** Error; Error Rates; Sampling Error; Variability, Measure of

**Further Readings**

Schmidt, F., & Hunter, J. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1*(2), 199–223.

Taylor, J. R. (1997). *An introduction to error analysis: The study of uncertainties in physical measurements.* Sausalito, CA: University Science Books.

Viswanathan, M. (2005). *Measurement error and research design.* Thousand Oaks, CA: Sage.

# Randomization Tests

Randomization tests use a large number of random permutations of given data to determine the probability that the actual empirical test result might have occurred by chance. Contrary to traditional parametric tests, they do not rely on the properties of known distribution, such as the Gaussian, for determining error probability, but construct the probability according to the actual distribution of the data. In parametric statistics, it is assumed that all potential sampled scores follow a theoretical distribution that is mathematically describable and well known. By transforming and locating an empirical score on the respective distribution, one can calculate a density function, which then gives the position of this score within this density function. This position can be interpreted as a $p$ value. Most researchers know that some restrictions apply for such parametric tests, so they may seek refuge in nonparametric statistics. However, nonparametrical tests are also normally approximated to a parametrical distribution, which is then used to calculate probabilities. In such a situation, strictly speaking, randomization tests apply.

Moreover, there are several situations in which parametric tests cannot be used. For instance, if a situation is so special that it is not reasonable to assume a particular underlying distribution, or if a single case research is done, other approaches are needed. In those, and in many more other cases, randomization tests are a good alternative.

## General Procedure

The basic reasoning behind randomization tests is quite simple and straightforward. Randomization tests assume that a researcher has done a controlled experiment, obtaining values for the experimental condition and values for the control condition, and in which an element of chance is involved as to either when the intervention started or which treatment a person was assigned to. Randomization tests also assume that a researcher has a range of possible values that could have popped up simply by chance. The researcher calculates test statistics, such as the difference between the experimental and control scores. The researcher then has the computer calculate all possible other differences (or does it by hand, in simple cases), obtaining a little universe of all possible scores. The researcher then counts the number of cases in which the empirically calculated difference is equal to or lesser or greater than all possible ones. This figure, divided by all possible scores, gives the researcher the true probability that the score obtained can be achieved just by chance.

The benefits of this procedure are as follows: It can be related directly to an empirical situation. It does not make any assumptions about distributions. It gives a true probability. It does not necessitate the sampling of many cases to reach a valid estimate of a sampling statistics, such as a mean or a variance (i.e., it can also be done with single cases). Most randomization tests are also associated with certain designs, but the basic reasoning is always the same.

## Practical Examples

Let us take an example for a single-case statistical approach, a randomized intervention design. The assumption of the intervention in this case is that a change in general level, say of depression or anxiety, is produced by the intervention. A researcher might want to think about such a design, if he or she has only a few patients with a particular problem and cannot wait until there are enough for a group study, or if he or she tailors interventions individually to patients, or if simply too little is known about an intervention.

In such a case, the researcher could do the following:

1. Decide on an outcome that can be measured daily or several times a day and that allows the researcher to draw conclusions about the state of the patient.
2. Think about how long it would take to decide whether the intervention was effective. This could be a week, a few days, or a month. This period determines the length of the baseline and follow-up observation period. During this period the researcher just measures to document the baseline condition and the condition after the intervention.
3. Between the baseline and the follow-up period, define a period of days within which the intervention can start. Also, consider how long the intervention should be applied and factor this into the planning. Let us assume, for the sake of the example, a period of 30 days between the baseline and the follow-up period.
4. Have the patient, or someone else, measure the desired outcome each day, or more often if possible and useful.
5. Decide at random the day of the variable period the intervention will start, and stick to the decision.
6. After the treatment, average all the scores from the start of the baseline to the start of the intervention, because this was the "real" baseline. Also, average all the scores from the day the intervention began until the end of the follow-up period. Calculate the difference. This will be the real, empirical difference.
7. Now repeat this procedure for every possible partitioning of time, first assuming the intervention had started at Day 1, calculating the difference between the baseline and the rest of the scores averaged over the rest of the period, then assuming it had started at Day 2, and so forth, until all the possible differences (in this example, 30 possible differences) are calculated. Here one can see why it is important to define a baseline and a follow-up period that is long enough: Simply by chance, it is possible that the intervention is to start on Day 1 of the variable period, and the researchers want a stable enough baseline to estimate the effect. Or it is possible that the intervention will be randomized to start only on Day 30. In that case the researcher will want a stable enough follow-up. If a longer treatment, say half a year, is expected until the treatment shows effects, the researcher could stipulate such a period between the end of the variable period, when the intervention can start, and the follow-up period.
8. Calculate all potential (in this case, 30) differences between baseline and follow-up, which in this example will give the researcher 30 statistical scores, only one of which is the real one, the one that was obtained by the actual intervention.
9. Now count how many of the possible difference scores are equal to or larger than the one obtained in the real experiment. Divide this number by the number of possibilities, in this case 30.
10. This gives the true probability that the score obtained was obtainable by chance alone. If none of all the other potential differences is larger than the score obtained, the probability is $p = 1/30 = .033$. Should the researcher find another score just as large as the empirically obtained one, the $p$ value would be $p = 2/30 = .066$.

This also shows that the $p$ value is related to two figures: One is the actual magnitude of the effect, that is, the effect size, or the power of the intervention to actually change the outcome parameter measured. This outcome is a matter of empirical reality. The other figure is the number of potential scores the researcher can calculate. In the example this was 30, because of the assumption that the intervention can start by chance on any one of 30 days. With that setup, a strong intervention whereby the researcher can be sure that it shifts

the outcome parameter sufficiently is needed. If the researcher is not quite sure how strong the intervention is, he or she might want to be on the safe side and increase the power of the test. This can be done either by increasing the measurement frequency to twice a day within the same time, yielding 60 potential differences, or by using a longer time. Similar to standard statistical testing, a researcher can manipulate the power of such a randomization test by increasing, in this case, not the number of participants, but the number of potential permutations, which in this design is related to the number of days or number of measurements.

Using such a design could potentially allow a researcher to make decisions on effectiveness of an intervention on an individual basis. It also allows the researcher to make decisions without having to withhold treatment, because there is no need for a control group. In this design, everyone is his or her own control, and the only restriction is that a participant might have to wait until the intervention begins. One could potentially also study a psychotherapeutic technique that takes longer to work. He or she would simply factor in the therapy time before the follow-up measurements.

Such an approach can also be used if one wants to study the short-term effects of interventions, similar to a crossover design in pharmacology. The same principle applies: The researcher defines possible periods, say days or weeks, within which the participant takes a substance or does something, perhaps watch a certain kind of video, and other days or weeks within which to take the control substance or watch other videos. The researcher measures the outcome of the intervention, and the time needed for measuring the outcome and for the effect to revert back to baseline is normally the time block. In the example case of pharmacology, this might be a week, or if the researcher wants to study the effect of funny versus violent movies, it might be a day. Here again, the kind of intervention and theoretically expected duration of effect define the length of the time block. The procedure is the following:

1. The researcher defines a number of blocks in which to study the experimental intervention and a number of blocks in which to apply the control condition. Here, as in the previous case, the number of possible time blocks defines the power of the test. Let us assume five blocks are used for the intervention and five blocks for the control, each block consisting of 1 day, making the whole experiment 10 days long.

2. The researcher randomizes the blocks, that is, decides at random on which day which intervention is going to be applied. In the extreme, it could be the first 5 days treatment, the next 5 days control, or the other way round. It could be alternating each day and any possible combination in between. In fact, there are then $10!/5!*5! = 252$ possibilities to permute $2*5$ days at random.

3. The researcher applies the interventions according to the design that has come out of the random choice.

4. The researcher measures the outcome after each intervention, depending on the intervention and the measure, perhaps even multiple times. Note that the length of the time blocks should be determined by such factual and operational criteria and define the design accordingly, not the other way round.

5. The researcher averages the outcomes for the control conditions and for the intervention conditions. Note that these could be averaged pre–post measurements per period, slopes of time curves in a continuous measurement of some physiological variable, or simply post measurements. This will yield one empirical difference between experimental conditions and control conditions.

6. The researcher now permutes the empirical results, under the assumption that some of the real experimental interventions had been controls, and the other way round. This means all possible differences between experimental and control conditions are calculated. Here is a simplified example:

> Let us assume a researcher had only three conditions each, the real sequence being a b b a a b, with "a" meaning treatment and "b" meaning control. Let the values of the outcome measurements associated with this design be 12 5 7 13 14 6. This would yield an average for the treatment of 13 and for the control of 6, with an empirical difference between the two conditions of 7. Now recalculate the values, permuting the design through all possible arrangements, which are $6!/3!*3! = 20$, assuming that the design had

been a a b a b b. This would yield an average of 10 for the treatment and 9 for the controls. The sequences are permuted and the averages and the differences recalculated accordingly:

| | |
|---|---|
| a a a b b b | $8 - 11 = -3$ |
| b b b a a a | $11 - 8 = 3$ |
| b a b a b a | $8 - 11 = -3$ |
| b b a b a a | $9 - 10 = -1$ |
| b a b b a a | $8.3 - 10.7 = -2.4$ |
| b a a b b a | $6 - 13 = -7$ |
| b a a a b b | $8.3 - 10.7 = -2.4$ |
| b b a a a b | $11.3 - 7.7 = 3.6$ |
| b b a a b a | $8.6 - 10.3 = -1.6$ |
| b a b a a b | $10.7 - 7.7 = 3$ |
| b a a b a b | $8.7 - 10.3 = -1.6$ |
| a a b b a b | $10.3 - 8.7 = 1.6$ |
| a b b b a a | $10.7 - 8.1 = 2.6$ |
| a b a a b b | $10.7 - 8.1 = 2.6$ |
| a b b a b a | $10.3 - 8.7 = 1.6$ |
| a a b a b b | $10 - 9 = 1$ |
| a a b b b a | $7.8 - 11.1 = -3.3$ |
| a b a b a b | $11 - 8 = 3$ |

One can see immediately that none of the permuted values is larger than the empirically obtained one of 7.

7. Now the researcher divides 1 by 20 (i.e., the number of potential permutations) which yields $p = .05$, conventional significance. This is the true probability of having obtained such a result purely by chance. One can also see that, had there been more possible permutations and only one value larger than every other possible one, the $p$ value would be smaller. To be able to obtain a value comparable to conventional significance, the minimum number of permutations possible has to be 20, yielding $p = 1/20 = .05$

Randomization tests can be employed whenever a researcher has single case designs in which some control is compared with some experimental intervention and an element of random shuffling can be introduced. All possible permutations of data have to be done and the calculation of the target statistics repeated for each possible permutation. If the effect is reasonably sized and the number of permutations is large enough, statistical decisions can be made on the basis of individual cases.

These individual $p$ values being independent can be agglutinated across independent cases using the Fisher-Pearson chi-square omnibus test $\chi^2 = -2*\sum\ln(p)$, with degrees of freedom $= 2*k$, where $k$ is the number of $p$ values to agglutinate, ln the natural logarithm, and $p$ the actual $p$ values. This yields a chi-square score that can be used to find the appropriate omnibus $p$ value.

In complex cases randomization statistics can be combined with Monte Carlo techniques. This would apply, for instance, if in an EEG experiment one would not want to fall back on parametric tests. Then a distribution can be generated out of the given data by random sampling of time points or time sequences. The general procedure is the same: The empirically found values are compared with the ones obtained by random sampling. The number of values equal to or larger than the empirically obtained ones is divided by the number of all possible values, yielding the true probability of having received the empirical result just by chance.

## Limitations

Randomization statistics are very robust because they do not depend on any assumptions. Their result is normally influenced by only two parameters: the real, empirically given effect and the number of possible permutations. The larger these possible permutations, the more powerful the test. However, increasing the number of possible permutations would have to be done at the cost of prolonging an experiment and increasing the number of measurements, making the experiment less feasible from a practical point of view.

However, where the expectation of finding a strong effect is realistic, such a design can produce a reliable result with very little effort. It is particularly feasible for single-case experiments, or in instances in which a large number of data sets have been acquired in comparatively few cases, such as in psychophysiological or psychological monitoring. In all these cases, randomization tests might be the only feasible way of reaching a scientifically sound conclusion.

Some caveats are in order: Randomization tests require certain conditions, such as a minimum number of possible, randomly permuted blocks across time or a minimum number of possible times when an intervention might begin. Such formal requirements might, in certain cases, conflict with clinical necessities. For instance, the effect of

a repeated intervention of, say, funny videos or stimulating pharmacological agents might taper off, creating a habituation and thus thinning out the effect. Or the assumption of a steady level of change of an intervention might not be warranted, or the measurement parameter used might suffer from a response shift, such that, over time, the inner standards of rating something changes. Thus, randomization tests and the corresponding designs have to be used with due care that the design be adapted to the research situation and question, and not the other way round.

Some also think that parametric statistics should be dumped altogether because the conditions under which they operate are violated more often than not, and it is rarely known how such a violation will affect the outcome of the test. Although this is probably true, the pragmatism of statistical testing has settled for the most part on using parametric tests because they are simple and straightforward to calculate and probably also because they have been incorporated in many computer programs and thus are user friendly. This should not prevent researchers, though, from using more adequate approaches, especially in conjunction with single-case designs.

*Harald Walach*

***See also*** Null Hypothesis; Random Assignment; Significance, Statistical

**Further Readings**

Edgington, E. S. (1995). *Randomization tests* (3rd ed.). New York: Dekker.

Onghena, P., & Edgington, E. S. (1994). Randomization tests for restricted alternating treatments designs. *Behaviour Research & Therapy, 32*(7), 783–786.

Onghena, P. (1994). *The power of randomization tests for single-case designs.* Unpublished doctoral dissertation. Katholieke Universiteit, Leuven, Belgium.

Wampold, B. E., & Furlong, M. J. (1981). Randomization tests in single-subject designs: Illustrative examples. *Journal of Behavioral Assessment, 3,* 329–341.

# Randomized Block Design

As enunciated by Ronald A. Fisher, a randomized block design (RBD) is the simplest design for comparative experiment using all three basic principles of experimental designs: randomization, replication, and local control. In this design, the treatments are allocated to the experimental units or plots in a random manner within homogeneous blocks or replications. It is appropriate only when the experimental material is heterogeneous. It may be used in single- or multifactor experiments. This entry considers the application, advantages and disadvantages, layout, randomization methods, and statistical analysis of randomized block designs.

## Applications

Randomized block design is most useful in situations in which the experimental material is heterogeneous and it is possible to divide the experimental material into homogeneous groups of *units* or *plots,* called *blocks* or *replications.*

It is mostly used in agricultural field experiments, where soil heterogeneity may be present due to soil fertility gradient, and in clinical trials on animals, where the animals (experimental units) vary in age, breed, initial body weight, and so on. A randomized block design can eliminate effects of heterogeneity in one direction only.

## Advantages and Disadvantages

The design is very flexible as no restrictions are placed on the number of treatments or replications. The statistical analysis is simple, and if some observations are missing or lost accidentally, one has recourse to the *missing plot technique.* The design is more efficient than completely randomized design as error variance is reduced as a result of blocking (assigning to blocks or replications), which increases the precision of the experiment. The design is not suitable if a large number of treatments are used or the experimental material is reasonably homogeneous.

## Layout of the Design

The plan of allocation of the treatments to the experimental material is called *layout of the design.* Let the $i$th $(i = 1, 2, \ldots, v)$ treatment be replicated $r$ times. Therefore $N = vr$ is the total number of required experimental units. Blocks or replications are formed perpendicular to the soil

**Table 1** Layout of Randomized Block Design in the Field

| | | | | |
|---|---|---|---|---|
| T1 | T5 | T2 | T4 | T3 |
| T5 | T4 | T3 | T2 | T1 |
| T3 | T2 | T1 | T5 | T4 |

fertility gradient in field experiments, or age, breed, initial body weight, and so on, in the clinical trials on animals, or education, age, religion, and so on, in social science experiments. Then, the treatments are allocated to the experimental units or plots in a random manner within each replication (or block). Each treatment has equal probability of allocation to an experimental unit within a replication or block.

Given in Table 1 is the layout plan of RBD in a field experiment with five treatments denoted by T1, T2, T3, T4, and T5, each replicated three times and allocated to 15 experimental units (rows are replications). The shaded areas are guard areas, nonexperimental areas, or areas under use as irrigation channels.

## Randomization

Some common methods of random allocation of treatments to the experimental units are illustrated in the following examples:

- For an experiment that has no more than 10 treatments, each treatment is allotted a number, using a 1-digit random number table. The researcher picks random numbers without replacement (i.e., a random number may not be repeated) from the random number table until the number of replications of that treatment is exhausted.
- For more than 10 treatments, researchers may use a 2-digit random number table or combine two rows or columns of 1-digit random numbers. Here each 2-digit random number is divided by the number of treatments, and the residual number is selected. The digit 00 is discarded.
- On small pieces of paper identical in shape and size the numbers 1, 2, . . . , N are marked. These are thoroughly mixed in a box, and the papers are drawn one by one. The numbers on selected papers are random numbers.
- The random numbers may also be generated by computational algorithms.

## Statistical Analysis

The data from RBD are treated by the method of analysis of variance for two-way classification. The analysis of data from RBD is performed using the linear fixed-effects model given as

$$y_{ij} = \mu + t_i + b_j + \varepsilon_{ij},\ i = 1, 2, \ldots, v;\ j = 1, 2, \ldots, r,$$

where $y_{ij}$ is the yield or response from the $i$th treatment in the $j$th replication, $t_i$ is the fixed effect of the $i$th treatment, and $\varepsilon_{ij}$ is the random error effect. Suppose one is interested in knowing whether wheat can be grown in a particular agroclimatic situation. One would randomly select some varieties of wheat from all possible varieties for testing in an experiment. Here the effect of treatment is random. When the varieties are the only ones to be tested in an experiment, the effect of varieties (treatments) is fixed.

The experimental error is a random variable with mean zero and constant variance ($\sigma^2$). These errors are assumed to be normally and independently distributed, with mean zero and constant variance for every treatment.

Let $N = vr$ be the total number of experimental units,

$$\sum_{i=1}^{v}\sum_{j=1}^{r} y_{ij} = N\bar{y}_{..} = G =$$

Grand total of all the observations,

$$\sum_{j=1}^{r} y_{ij} = r\bar{y}_{i.} = T_{i.} =$$

Total response from the experimental units receiving the $i$th treatment, and

$$\sum_{i=1}^{v} y_{ij} = v\bar{y}_{.j} = T_{.j} =$$

Total response from the experimental units in the $j$th block then

**Table 2** Analysis of Variance (ANOVA) for Randomized Block Design

| *Source of Variation* | *Degrees of Freedom* | *Sum of Squares* (SS) | *Mean Sum of Squares* | *Variance Ratio* (F) |
|---|---|---|---|---|
| Treatments | $v-1$ | $TrSS$ | $s_t^2 = TrSS/(v-1)$ | $F_t = s_t^2/s_e^2$ |
| Blocks | $r-1$ | $RSS$ | $s_b^2 = RSS/(r-1)$ | $F_t = s_b^2/s_e^2$ |
| Error | $(v-1)(r-1)$ | $ErSS$ | $s_e = ErSS/(N-v)$ | |
| Total | $N-1$ | $TSS$ | | |

*Note:* $TSS$ = total sum of squares; $TrSS$ = treatments sum of squares; $RSS$ = block sum of squares; $ErSS$ = error sum of squares.

$$\sum_{i=1}^{v}\sum_{j=1}^{r}(y_{ij}-\overline{y}_{..})^2 = r\sum_{i=1}^{v}(\overline{y}_{i.}-\overline{y}_{..})^2 + v\sum_{j=1}^{r}(\overline{y}_{.j}-\overline{y}_{..})^2 + \sum_{i=1}^{v}\sum_{j=1}^{r}(y_{ij}-\overline{y}_{i.}-\overline{y}_{.j}+\overline{y}_{..})^2.$$

That is, $TSS = TrSS + RSS + ErSS$ where $TSS$, $TrSS$, $RSS$, and $ErSS$ stand for total, treatments, block, and error sum of squares. This implies that the total variation can be partitioned into three components: variation due to treatments, blocks, and experimental error. Table 2 shows an analysis of variance table for randomized block design.

Under the null hypothesis, $H_0 = t_1 = t_2 = \cdots = t_v$, that is, all treatment means are equal, the statistic $F_t = s_t^2/s_e^2$ follows the $F$ distribution with $(v-1)$ and $(v-1)(r-1)$ degrees of freedom.

If $F_t$ (observed) $\geq F$ (expected) for the same degrees of freedom and at a specified level of significance, say $\alpha$%, then the null hypothesis of equality of treatment means is rejected at $\alpha$% level of significance, and $F_t$ is said to be significant at $\alpha$% level of significance.

If $H_0$ is rejected, then the pairwise comparison between the treatment means is made by using the critical difference test statistic,

$$t = \frac{\overline{y}_{i.}-\overline{y}_{j.}}{s_e\sqrt{\frac{2}{r}}}.$$

The critical point for the two-tailed test at $\alpha$ level of significance is $t_{(v-1)(r-1),1-\alpha/2}$. If the test statistic is more than or equal to the critical point, the treatment means $i$ and $j$ are said to differ significantly.

Under the null hypothesis, $H_0 = b_1 = b_2 = \cdots = b_r$, that is, all block means are equal, the statistic $F_b = s_b^2/s_e^2$ follows the $F$ distribution with $(r-1)$ and $(v-1)(r-1)$ degrees of freedom.

If $F_b$ (observed) $\geq F$ (expected) for the same degrees of freedom and at a specified level of significance, say $\alpha$%, then the null hypothesis of equality of block means is rejected at $\alpha$% level of significance.

If $H_0$ is rejected, then the pairwise comparison between the block means is made by using the critical difference test statistic,

$$t = \frac{\overline{y}_{.i}-\overline{y}_{.j}}{s_e\sqrt{\frac{2}{v}}}.$$

The critical point for the two-tailed test at $\alpha$ level of significance is $t_{(v-1)(r-1),1-\alpha/2}$. If the test statistic is more than or equal to the critical point, the block means $i$ and $j$ are said to differ significantly.

If the null hypothesis, that is, equality of means for blocks, is not rejected, then one may infer that either the blocking has not been done properly or it is not feasible or the experimental material is reasonably homogeneous. The critical difference is also called *least significant difference*.

### *Coefficient of Variation*

The coefficient of variation (CV) is a relative measure of experimental error. It is given by CV% = $(s_e/\text{GM}) \times 100$, where $s_e^2$ and GM stand for error variance and grand mean, respectively. The coefficient of variation empirically lies between 10% and 20% for yields from field experiments. However, it may be even less than 5% for characters showing little variability, such as oil content, protein content, and so on. The researcher has prior information about the CV% from similar experiments. If the CV% is very high, the researcher may suspect that the uniform application of agronomical practices and statistical

methods has not been carried out properly and may reject the experimental findings. If the CV% is very low, the genuineness of the data is suspect and suggests proper examination of the data.

### *Missing Plot Technique*

Sometimes a treatment value may be missing for accidental reasons and not as a result of the effect of treatment. In such cases, these values can be estimated and statistical analysis carried out after substituting the estimated value. The formula for one missing value is given by

$$X = (Tv - Rr - G)/(v - 1)(r - 1),$$

where $X$ is the estimated value for the missing plot, $v$ is number of treatments, $r$ is number of replications, $T$ is total of the treatment that contains the missing value, $R$ is total of the block in which the missing value occurs, and $G$ is total of all observations excluding the missing value.

The estimated value is substituted, and then the analysis is done as usual. Since one observation is lost, the error degrees of freedom and total degrees of freedom should be reduced by one. The standard error of difference between the mean of the treatment with a missing value and the mean of any other treatment is given by

$$SE_d = \sqrt{s_e^2}\{(2/r) + v/(r - 1)(v - 1)\},$$

where $s_e^2$ is error variance in the analysis of variance.

*Kishore Sinha*

***See also*** Analysis of Variance (ANOVA); Completely Randomized Design; Experimental Design; Fixed-Effects Models; Research Design Principles

**Further Readings**

Bailey, R. A. (2008). *Design of comparative experiments*. Cambridge, UK: Cambridge University Press.
Box, G. E. P., Hunter, W. G., & Hunter, J. S. (1978). *Statistics for experiments*. New York: Wiley.
Dean, A., & Voss, D. (1999). *Design and analysis of experiments*. New York: Springer-Verlag.
Montgomery, D. C. (2001). *Design and analysis of experiments*. New York: Wiley.

## Random Sampling

In random sampling, the sample is drawn according to prespecified chances from the population, and thus it is also called *probability sampling*. Since planned randomness is built into the sampling design according to the probabilities, one can use these probabilities to make inferences about the population. For example, if one uses the sample mean to estimate the population mean, it is important to know how the sample is being drawn since the inference procedures such as confidence intervals will depend on the sampling scheme. Similarly, hypothesis testing on the population mean also depends on the sampling scheme used.

One important purpose of random sampling is to draw inferences about the population. On the other hand, if the sampling scheme is nonrandom, that is, not all outcomes have a known chance of occurring, the sample is likely to be biased. It is difficult to draw inferences about the population on the basis of nonrandom sampling. Some common nonrandom sampling methods are *quota sampling*, *convenience sampling*, and *volunteer sampling*.

The following sections provide a brief outline of some of the most commonly used random sampling schemes: simple random sampling, systematic sampling, stratified sampling, and cluster sampling.

### Simple Random Sampling

Simple random sampling is the simplest random sampling scheme. There are two types of simple random sampling: with or without replacement. For simple random sampling without replacement, ($n$ distinct units are selected from a population of $N$ units so that every possible sample has the same chance of being selected). Thus the probability of selecting any individual sample $s$ of $n$ units is

$$p(s) = \frac{1}{\binom{N}{n}} \quad \text{where} \quad \binom{N}{n} = \frac{n!(N-n)!}{N!}.$$

Sampling without replacement is often preferred because it provides an estimate with smaller variance. However, even though for a finite population, when a sample has been selected, it does not

provide additional information to be included in the sample again. One practical advantage to sample with replacement is there that is no need to determine whether a unit has already been sampled. In addition, when sampling with replacement, each drawing of each sample is independent, and the probability formulas for the samples are usually easier to derive. For simple random sample with replacement, a unit may be selected more than once, and each draw of a unit is independent: Each unit in the population has the same probability of being included in the sample. When the sample size is much smaller than the population size, the two sampling schemes are about the same. Otherwise, it is important to distinguish whether the simple random sampling is with or without replacement since the variance for statistics using these two schemes is different, and thus most inference procedures for these two schemes will be different. For example, in estimating the population mean under simple random sampling without replacement, the estimated variance for the sample mean is

$$\hat{V}(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{s^2}{n}$$

where $s^2$ is the sample variance.

Note that the factor

$$\left(1 - \frac{n}{N}\right)$$

is called the *finite population correction factor*. Under simple random sampling with replacement, there is no need for this factor, and the estimated variance for the sample mean is simply

$$\hat{V}(\bar{y}) = \frac{s^2}{n}.$$

In designing a survey, one important question is how many to sample. The answer depends on the inference question one wants to answer. If one wants to obtain a confidence interval for the parameter of interest, then one can specify the width and level of significance of the confidence interval. Thompson (2002), chapter 4, provides an excellent discussion for determining sample sizes when one is using a simple random sample.

One advantage of simple random sampling is that making inferences is simple and easy with this sampling scheme. It is easy to apply to small populations. With a large population, it is often difficult for researchers to list all items before they draw randomly from the list. This difficulty limits the use of simple random sampling with large populations.

## Systematic Sampling

Systematic sampling is easier to perform in the field and provides a useful alternative to simple random sampling. The researcher starts at a random point and selects items that are a fixed distance apart. When no list of the population exists or if a list is in approximately random order, systematic sampling is often used as a proxy for simple random sampling. In cases in which the population is listed in a monotone order, systematic sampling usually results in estimators with smaller (though sometimes unestimable) variances than those of simple random sampling. Repeated systematic sampling consists of more than one systematic sample, each with a different random starting point. Using the variability in the subsample means, one can get a measure of the variance of that estimate in the entire sample. When the population is in periodic order, then it is important to choose the period appropriately in order to get a representative sample.

One can classify systematic sampling as a special case of cluster sampling. The population is partitioned into clusters such that the clusters consist of units that are not contiguous. For systematic sampling to be effective, the ideal cluster should contain the full diversity of the population and thus be representative of the population. One simple example for systematic sampling is to interview subjects in a long queue about their political affiliation. If there are 700 people in the queue and one plans to take a sample of size 50, then to take a systematic sample, one chooses a number randomly between 1 and 14 and then every 14th element from that number on. If the people form the queue in a random order, then systematic sampling will produce the same result as a simple random sample. If people with similar political affiliation stay close together in the queue, then systematic

sampling will likely be more effective than simple random sampling.

Systematic sampling is often used in industry for quality control purposes. When there is a malfunction beginning with and continuing from a certain item, a systematic sample will provide a representative sample of the population whereas simple random sampling may over- or underrepresent the defective items.

## Stratified Sampling

When the population is partitioned into strata and a sample is selected from each stratum, then the sampling scheme is called *stratified sampling*. One important characteristic of the strata is that the elements in the population belong to one and only one stratum. Stratified random sampling refers to the situation in which the sample selected from each stratum is selected by simple random sampling. Alternatively, one may select a sample from each stratum by systematic sampling or other schemes. When the stratum consists of units more similar in the variable of interest, stratified sampling usually results in estimates with smaller variances. One example of stratified sampling is a study to estimate the average number of hours per week that undergraduates from a large public university spend studying. One can stratify by department and conduct stratified sampling. If the time that students within a department spend studying is about the same whereas there is large variability between departments, then a stratified sample will have smaller variance than a random sample will. On the other hand, if one wants to estimate the average height of undergraduate students in that university, one may choose to use stratified random sampling by department for ease of administration. Note that the resulting estimate from stratified sampling may not have the benefit of smaller variances than simple random sampling would offer because students in the same department may be of very different height.

One important practical consideration in stratified sampling is how many samples to allocate to each stratum. When one has no prior information about the variances of the strata, then the optimal allocation is to assign sample sizes proportional to stratum size. If the sampling cost is the same for each stratum and the standard deviations of each stratum can be estimated, then the optimal allocation is to assign sample sizes proportional to the product of stratum size and stratum standard deviation. That is, the optimal allocation assigns larger sample sizes to larger or more variable strata.

Quite often, the characteristic that one wants to base the stratification on may not be known before the survey is carried out. For example, we may want to stratify voters by gender, but gender information may be known only after the voter is contacted. *Double sampling* (also called *two-phase sampling*) may be used in such situations. In double sampling, a simple random sample is collected from the population for the purpose of classifying these items to the appropriate stratum. A second sample is then selected by stratified random sampling from the first sample. One important use of double sampling is to adjust for nonresponse in surveys. In addition to the responses they obtain, the researchers stratify the population into response and nonresponse groups according to a first sample of the population. To obtain responses, the researchers re-calls people from the nonresponse group and offer them more incentives to respond. See Scheaffer, Mendenhall, and Ott (1995), Chapter 11.5, for guidelines for determining the number of callbacks.

## Cluster Sampling

Sometimes listing the elements of a population is costly or not available, but a listing of clusters is available. Also, the cost of sampling elements within a cluster is often less than the cost of other sampling methods. In such cases, cluster sampling is a good choice. Cluster sampling and stratified sampling differ in that in stratified sampling, a simple random sample is taken from every stratum, whereas in cluster sampling, a simple random sampling of the clusters is taken. The effectiveness of cluster sampling depends on the variances from using the sampled clusters and the costs of sampling these clusters and the units within the clusters. Quite often, cluster sampling has less precision than simple random sampling from the same universe because units within the same cluster are usually more similar and thus provide less information per unit. In cluster sampling (or *one-stage cluster sampling*), all units of a cluster are sampled if the cluster is a sample. If, after selecting a sample of clusters, a sample of units within the clusters is

selected, the design is called *two-stage cluster sampling*. Similarly, *multistage cluster sampling* may be carried out. Suppose one wants to estimate the average hours per week undergraduates from a large public university spend studying. One can treat each university department as a cluster. To conduct one-stage cluster sampling, one randomly selects departments in that university, then samples all the undergraduate students in those departments. If instead one samples some individuals from the selected departments, one is using two-stage cluster sampling. An example for multistage sampling is to first select states, then universities within the states, then departments within the universities, and then students within the departments.

In two-stage cluster sampling, various designs exist for selecting the clusters in the first stage. Frequently one uses probability proportional to size, in which the probability of a cluster's being selected is proportional to the number of units it contains. Other sampling schemes commonly employed at each stage include simple random sampling and systematic sampling.

When designing a cluster sample, one decides what overall precision is needed, what size the clusters should be, and how many to sample from each cluster. Lohr (1999) Chapter 5 provides a detailed discussion.

### Conclusion

For a complex survey, rarely will only one type of sampling scheme be used. Often many sampling schemes will be combined. For example, a population may be divided into strata, and multistage cluster sampling can be used to sample from each stratum. Simple random sampling may be used in one stage whereas systematic sampling be used in the next stage. The four basic random sampling schemes discussed here have important implications for the estimators to use and the inference procedure for each estimator.

*Mosuk Chow*

***See also*** Population; Sample; Sampling

#### Further Readings

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: Wiley.

Kish, L. (1965). *Survey sampling*. New York: Wiley.

Lohr, S. L. (1999). *Sampling: Design and analysis*. Pacific Grove, CA: Duxbury.

Scheaffer, R. L., Mendenhall, W., & Ott, R. L. (1995). *Elementary survey sampling* (5th ed.). Pacific Grove, CA: Duxbury.

Thompson, S. K. (2002). *Sampling* (2nd ed.). New York: Wiley.

## Random Selection

Random selection is a precise, scientific procedure whereby each unit in a population has an equal chance of selection for inclusion in a sample. This concept underlies the premise of probability sampling and is central to probability methodologies and generalizability. Random selection eliminates sampling selection bias but introduces random error in its place. Random selection is the only valid strategy for obtaining a representative sample in research.

### Random Selection and Probability Sampling

*Probability theory,* which emerged in response to a desire to better understand card games and gambling, is the branch of mathematics dealing with how to estimate the chance that events will occur. In probability theory, the probability of something occurring is usually expressed as the ratio between the number of ways an event can happen and the total number of things that can happen. For example, the probability of picking a heart from a deck of cards is 13/52. In other words, inferences are made about events based on knowledge of the population.

There are two types of sampling methodologies that can be used in research: *probability* and *nonprobability*. Each methodology emphasizes different sample selection procedures. Probability sampling methods all use a random selection procedure to ensure that no systematic bias occurs in the selection of elements. The term *random selection* is commonly used in the context of experimental research, while the term *random sampling* is more common to survey research; however, the underlying principle is the same. In probability methodologies, each unit of the population has a known nonzero chance of being selected.

Probability sampling is typically used for quantitative research and large sample sizes and is considered to be representative of reality by statistical inference.

The most simple and common example of random selection is the tossing of a perfect coin. Each toss of the coin can result in two outcomes, heads or tails, with both outcomes having an equal probability of occurrence, that is, 50%. Moreover, each event is totally independent of previous selections and is also mutually exclusive, that is, a head and a tail cannot occur at the same time.

There are several types of probability sampling methods, namely, simple random sampling, systematic sampling, stratified random sampling, and cluster sampling.

### *Simple Random Sampling*

In simple random sampling, every element in the population has an equal chance of being selected. This method is ideal for drawing a sample but would be quite time-consuming and difficult if a large population is involved. The basic procedure involves the following: identifying the research population, enumerating each element, and devising a selection method to ensure that each element has an equal chance of selection. For example, if a class has 100 students and the teacher wanted to take 10 students on a field trip, the teacher could write each student's name on a piece of paper, place all the names in a bottle, and then randomly select 10 names. A more sophisticated alternative would be to use a random number table to select the 10 students. Using this technique, each student would be assigned a number from 1 to 100. An arbitrarily starting point on the random number table would then be selected. The teacher would then move down the columns searching for the first 10 numbers on the table that carry 1 to 100 as their first 3 digits.

### *Systematic Sampling*

Systematic sampling is a variant of simple random sampling whereby every *n*th element is selected from the research population list or sampling frame. The procedure involves identifying the research population, determining the interval for every *n*th element (population divided by desired sample size, e.g., 1,000/100 = 10), selecting the first element randomly (this element should fall within the sampling interval, e.g., 1 to 10), and finally selecting every *n*th element until the sample size is achieved. Thus, if the first element is 7, the sample would be comprised of the numbers 7, 17, 27, 37, 47, 57, 67, 77, 87, and 97.

### *Stratified Random Sampling*

Stratified random sampling involves dividing the population into subgroups or strata and taking a random sample from each stratum through either simple random or systematic sampling. The advantage of this method is that key subgroups, as well as the overall population, are accurately represented. Populations can be stratified by gender, age, location, size, and so on. The criterion for stratification is determined by the research question or hypothesis. This method is suitable only for fairly large populations.

### *Cluster Sampling*

Cluster sampling is a process in which a random selection of naturally occurring clusters (e.g., schools for sampling students, blocks for city residents, countries for general population) is first performed, followed by a random selection of elements within the clusters. This methodology is best suited for large populations scattered across a wide geographical area, where an exhaustive listing of all the elements would be an unreasonable proposition and prohibitively expensive.

## Benefits

There are two primary benefits to random selection. First, random selection removes the danger of sample selection bias caused by drawing a nonrepresentative sample and associated with nonprobability sampling methods. Second, random selection (a) allows the researcher to estimate sampling error that causes differences between the sample and the population and (b) permits *statistical inference*. Statistical inference refers to the statistical processes, namely, confidence intervals and significance tests, used to make inferences concerning some unknown aspect of a population. More specifically, statistical inference indicates the confidence researchers can attach to their findings when

they declare that the findings represent good estimates of what exists in the population.

## Random Selection and Sampling

Researchers are generally interested in learning something about large groups or things. Thus, the major questions driving sampling are, What do you want to know? To whom do you want to generalize? There are two types of populations that a researcher can sample—*theoretical* and *accessible*. In practice, researchers sample an accessible population because a sample of a theoretical population would be quite an arduous undertaking and the cost of such a sample would be quite prohibitive. In addition, trying to obtain a sample from a theoretical population would be quite difficult because no accurate listing of the population may exist. For example, for a study of homeless urban teenagers in the United States, the researcher might use an accessible population comprising five urban areas in the United States.

Samples therefore offer a practical solution to the study of populations. *Sampling* may thus be defined as the process of selecting units from a population of interest. It is a key component of sound research design because the manner in which a sample is selected will determine whether generalizations can be made from the sample to the research population. In other words, only probability methodologies are suitable for making generalizations because these are the only methodologies that are based on random selection.

Researchers who wish to generalize must follow very stringent rules for sampling and sample size. Generally, the larger the population, the smaller the sample ratio that is needed to obtain a representative sample. Doubling a sample size does not double accuracy; however, it does double cost. Standard sampling ratios are 500 (50%), 1,000 (30%), 10,000 (10%), 150,000 (1%), and 1,000,000 (.025%). Today, sample size calculators are available on the Web that can generate sample size calculations in mere seconds.

## Sampling Frame

To facilitate probability sampling, researchers must work with a sampling *frame*, that is, an exhaustive list of all the units that comprise the research population. The sampling frame is critical to probability sampling. In order to construct or select a sampling frame, a researcher must have a very clear idea of the kinds of data sources that will be required for the particular study. Common examples of sampling frames include telephone directories, college enrollment records, patient records, membership lists, business directories, universities, cities, academic journals, unions, and professional associations, to name a few.

## Generalization

Generalization refers to the extent to which researchers can make statistical inferences from a sample to a population, based on information obtained from a sample. Generalizations should be made cautiously—generalizations from a particular study should be taken as representing only the sum of elements from the initial sampling frame. Moreover, in order to make generalizations, it is critical that the initial sampling frame be relatively complete, with all elements having equal representation. In addition, the sample size must meet the statistical requirements necessarily for generalizability. For example, if a study was being done on the qualifications and evaluation experience of members of the American Evaluation Association (AEA), which has a diverse international membership, but the list from which the random sample was extracted included only members of the association who reside in the United States, it would be incorrect to use this study to make a generalization about the qualifications or experience of all AEA members. Likewise, even if the list included every member, but a random sample of only 10 individuals was taken, these results could not be generalized because a sample of 10 is inadequate for making a generalization.

## Computer Software for Generating Random Numbers

Prior to the technological era, random section was a cumbersome, tedious, manual process, particularly where large samples were involved. Selection was generally accomplished by using a random number table, that is, a table containing lists of numbers ordered solely on the basis of chance.

These tables are generally located at the back of most social science research texts. Techniques such as randomly picking 100 entries from a hat, or every *n*th item to obtain a sample of 100, were not at all uncommon.

Fortunately, with the advent of modern technology, this process is now remarkably simple. Modern computer programs (e.g., EXCEL; SAS; SPSS, an IBM company formerly called PASW (R) Statistics; SYSTAT; KWIKSTAT; STATMOST) now have built-in random number generators that can generate a random sample of any size in mere seconds. However, for the computer to generate this list, all elements in the sampling frame must be numbered. This task can also be performed in mere seconds. In addition, techniques such as *random digit dialing* can greatly simplify the process of random selection when telephone surveys are being conducted. This process is facilitated by an automated procedure whereby a machine dials random numbers within designated phone prefixes. However, since only approximately 20% of the numbers turn out to be working residential numbers, many telephone calls are required with this procedure in order to obtain the desired sample. Notwithstanding, random digit dialing is still quite a popular procedure for conducting polls. It is used by many large media organizations such as the New York Times and CBS News. A commonly used and more reliable variant of random digit dialing is a technique referred to as the *Waksberg method*. With this technique approximately two thirds of all calls are placed to working resident telephones.

Today, several firms (e.g., Survey Sampling, Inc., Genesys Sampling Systems) specialize in providing samples for telephone surveys. These firms provide telephone lists that are scientifically sound, with business numbers excluded. Such firms are in great demand by organizations who conduct survey research via telephone interviewing.

*Nadini Persaud*

***See also*** Experimental Design; Probability Sampling; Probability, Laws of; Quasi-Experimental Design; Random Sampling; Sample; Sample Size

**Further Readings**

Babbie, E. (2004). *The practice of social research* (10th ed.). Belmont, CA: Wadsworth/Thomson Learning.

Bouma, G. D., & Atkinson, G. B. J. (1995). *Social science research; A comprehensive and practical guide for students*. Oxford, UK: Oxford University Press.

Ruane, J. M. (2005). *Essentials of research methods: A guide to social sciences research*. Oxford, UK: Blackwell.

Schutt, R. K. (2001). *Investigating the social world: The process and practice of research* (3rd ed.). Thousand Oaks, CA: Pine Forge Press.

Vogt, P. W. (1999). *Dictionary of statistics and methodology: A non-technical guide for the social sciences*. Thousand Oaks, CA: Sage.

Weisberg, H. F., Krosnick, J. A., & Bowen, B. D. (1996). *An introduction to survey, research, polling and data analysis* (3rd ed.). Thousand Oaks, CA: Sage.

## Random Variable

A random variable, as defined by J. Susan Milton and Jesse Arnold, is an assigned value, usually a real number, to all individual possible outcomes of an experiment. This mapping from outcome to value is a one-to-one relationship, meaning that for every experimental outcome, one and only numeric value is associated with it.

Take for example the tossing of two coins. The outcomes of such an experiment will have the following possibilities: heads, heads; heads, tails; tails, heads; and tails, tails. For this experiment, the random variable will map those categorical outcomes to numerical values, such as by letting the random variable $X$ be the number of heads observed. Preserving the above order, the possible outcomes for $X$ are 2, 1, 1, and 0. This particular set lists each element with an equal probability; in this case, the four outcomes each have a 25% chance of occurring.

The listing of all possible outcomes of an experiment is called the *sample space*. In the case of the random variable $X$, which counts the number of heads, the sample space can be simplified to the set (2, 1, 0) where the values 2 and 0 each have a 25% chance of occurring and the value 1 has a 50% chance, since the outcomes (heads, tails) and (tails, heads) both have a value of 1 in terms of $X$.

A random variable is the unpredictable occurrence of a subset of outcomes from the aforementioned sample space. There are two types of

random variables, namely, *qualitative* and *quantitative,* with quantitative random variables being the more widely used version in both applied and theoretical settings.

Qualitative random variables' defining characteristic is that they can be categorized into specific groups. The focus of the variable is on a "quality" that cannot be mathematically measured. Some examples of qualitative variables are characteristics such as hair color, eye color, or agreeability toward statistics. These variables contain outcomes that are not arithmetically comparable to one another, and they are also referred to as *nominal variables*.

Quantitative random variables are assigned an actual numeric value for each element that occurs in the sample space, as in the coin-tossing example above. A prerequisite for quantitative variables is that the numeric values will have a mathematical relationship with other values. This relationship may be as rudimentary as one value's being greater than another. Examples of quantitative variables include a person's height, age, or the position in which they finished a race. Quantitative variables can be further classified into *interval, ratio,* and *ordinal* scales.

In an interval scale the key principle is that the distance between any value and its neighboring value will be the same distance found between any other two adjacent values on the scale. This scale does not contain an absolute zero point, so it does not make sense to construct a ratio of interval values. For example, the temperature scale of Fahrenheit is an interval scale, and it is not proper to say that a 100-degree day is twice as hot as a day with a high of 50 degrees Fahrenheit.

In contrast to the interval scale, the ratio scale contains a true zero point. The existence of an absolute zero allows for an understandable ratio of any two values on the scale. A corollary to the previous statement is that a distance between any two points anywhere along the scale continuum has the same meaning. An example of a ratio scale is a measurement made in centimeters. The difference between 10 and 20 centimeters is the same as that found between 132 and 142 centimeters. Also, it is an obvious assumption that an object that is 10 centimeters long is twice as long as an object 5 centimeters in length.

The ordinal scale provides information about subjects relative to one another. The difference between any two subjects is indeterminable because the values assigned in the ordinal scale are not based on magnitude of an attribute. A 5-point Likert-type satisfaction scale, in which 5 is *very satisfied* and 1 is *very dissatisfied,* is an example of an ordinal scale. That is, an answer of 4 indicates that a respondent is more satisfied than a person who answered 2. Yet, the person answering 4 is not twice as satisfied as the person who answered with a satisfaction score of 2.

Within statistics, random variables are involved in three fundamental areas, namely *cumulative distributions, probability distributions,* and *expected values,* under the condition that the random variable in question, whether it is discrete or continuous, can be modeled by some formula with respect to probability.

The *cumulative probability* is the chance of finding some value, $a$, or any other value in the sample space that is less than $a$. The value of $a$ must be greater than the smallest value, or the lower bound, of the sample space.

For the discrete case where $b < x < c$,

$$P(X \leq a) = \sum_{b}^{a} x_i.$$

For the continuous case where $b < x < c$,

$$P(X \leq a) = \int_{b}^{a} f(x)dx.$$

A probability distribution defines the probability of a random variable's being observed within a set of given values for either the discrete or continuous case. In particular, the probability distribution is the first-order derivative of the cumulative probability distribution. A side note is that the probability of a single random variable in the discrete case exists, but not in the continuous case.

For the discrete case of a probability for the $d$th and $e$th values and in between,

$$P(d \leq x \leq e) = \sum_{d}^{e} x_i.$$

For the continuous case of a probability between the $d$th and $e$th values,

$$P(d \le x \le e) = \int_d^e f(x)dx.$$

The expected value is the mathematical expectation of the random variable. This value is sometimes called the mean. The expected value may or may not actually be an observable value in the sample space. Theoretically, the expected value is the value that would be observed if an infinite number of experiments were conducted and the outcomes were all averaged. The expected value is denoted as E(x) and is found by multiplying each element in the sample space by its associated probability and either summing or integrating those products across the entire range for which the random variable is defined.

For the discrete case where $b < x < c$,

$$P(X \le a) = \sum_b^a x_i p(x_i).$$

For the continuous case where $b < x < c$,

$$P(X \le a) = \int_b^a xf(x)dx.$$

*Michael T. McGill*

***See also*** Probability, Laws of; Variable

**Further Readings**

Bronshtein, I. N., & Semendyayev, B. (1997). *Handbook of mathematics* (K. A. Hirsch, Trans.). Berlin, Germany: Springer.

Milton, J. S., & Arnold, J. C. (1990). *Introduction to probability and statistics: Principles and applications for engineering and the computing sciences*. New York: McGraw-Hill.

Ott, R. L. (1993). *An introduction to statistical methods and data analysis*. Belmont, CA: Duxbury.

# RANGE

The range of a distribution of values is the difference between the highest and lowest values of a variable or score. In other words, it is a single value obtained by subtracting the lowest (minimum) from the highest (maximum) value. Range provides an indication of statistical dispersion around the central tendency or the degree of spread in the data. There are several methods to indicate range, but most often it is reported as a single number, a difference score. However, in some contexts, the minimum and maximum are both presented to reflect the range as a pair of values. For example, if the data included values from 10 through 42, the range may be noted as either 32 (the difference between the highest and lowest value; i.e., 42 − 10) or the pair of numbers reflecting the lowest and highest values (10, 42).

## Examples of Range

The range is the smallest number subtracted from the largest number.

*Example 1:* What is the range of the following data set (1, 2, 3, 4, 2, 3, 4, 5, 5, 5)?

Range is $5 - 1 = 4$; otherwise noted as $(1, 5)$.

*Example 2:* What is the range of the following data set (1, 1, 1, 1, 1, 1, 1, 1, 1, 5)?

Range is $5 - 1 = 4$.

*Example 3:* What is the range of the following data set (1, 1, 1, 2, 2, 2, 2, 2, 10)?

Range is $10 - 1 = 9$.

Range is typically used to characterize data spread. However, since it uses only two observations from the data, it is a poor measure of data dispersion, except when the sample size is large. Note that the range of Examples 1 and 2 earlier are both equal to 4. However, the two data sets contain different frequencies of each value. It is an unstable marker of variability because it can be highly influenced by a single outlier value, such as the value 10 in Example 3. One way to account for outliers is to artificially restrict the data range. A common method is to adjust the ends or extremes of the distribution curve. Typically, *interquartile range* (i.e., the distance between the 25th and 75th percentiles) is used. By definition, this contains 50% of the data points in a normally distributed data set. In some contexts, it may be appropriate to use the *semi-interquartile range*, which covers 25% of the data points and is even

less subject to variation due to scatter. Likewise, *interdecile range* evaluates only the 1st and 9th deciles or spread between 10th and 90th percentiles. This is similar to the median concept.

### Examples of Range Restriction

In correlational studies (examining how one variable relates to another), the sampling procedure, or the measuring instrument itself, sometimes leads to a restriction in the range of either or both of the variables being compared. The following examples illustrate the point.

1. The height of professional basketball players is more restricted than height in the general population.
2. The IQ scores of college seniors are more restricted than the IQ scores of the general population.
3. The vocabulary test scores of fifth graders are more restricted than the vocabulary test scores of all grade school children.
4. Tests that are extremely easy or difficult will restrict the range of scores relative to a test of intermediate difficulty. For example, if the test is so easy that everyone scores above 75%, the range of scores will obviously be less than a test of moderate difficulty in which the scores range from 30% to 95%.

### Statistical Implications

Data that are restricted in range may produce a threat to interpretation of outcomes because they artificially reduce the size of the correlation coefficient. This is because restriction of range would be reflected in a smaller variance and standard deviation of one or both marginal distributions. For example, consider the use of height of basketball players to predict performance, as measured by points scored. Who is likely to score more, a person 62 inches (157.5 cm) tall or a person 82 inches (208.3 cm) tall? Nothing is certain, but a person who is 82 inches tall seems a better bet than someone 62 inches tall. Compare this situation with the question, Who is likely to score more, a person 81 inches (205.7 cm) tall or a person 82 inches tall? Here the prediction is much less certain. Predicting from the narrow range of 1 inch (2.54 cm) is extremely difficult compared with predicting from a range of 20 inches (50.8 cm). In sum, range restriction likely results in smaller variance, which in turn reduces the correlation coefficient.

### Other References to Range

Other references to range include the following:

*Crude range:* the difference between the highest and lowest score, another term commonly used for range.

*Potential crude range:* a potential maximum or minimum that will emanate from a measurement scale.

*Observed crude range:* the actual smallest and greatest observation that resulted from a measurement scale.

*Midrange:* the point halfway between the two extremes. It is an indicator of the central tendency of the data. Much like the range, it is not robust with small sample size.

*Studentized range distribution:* range has been adjusted by dividing it by an estimate of a population standard deviation.

*Hoa T. Vo and Lisa M. James*

***See also*** Central Tendency, Measures of; Coefficients of Correlation, Alienation, and Determination; Correlation; Descriptive Statistics; Sample Size; "Validity"

#### Further Readings

Lockhart, B. S. (1997). *Introduction to statistics and data analysis for the behavioral sciences*. New York: Macmillan.

## RATING

Ratings are summaries of behaviors or attitudes that have been organized according to a set of rules. Although the general use of the term in research can refer to written data, such as an anecdotal report, a *rating* usually refers to a number assigned according to a defined scale.

Ratings can be *quantified,* as numbers or scores chosen to identify the position on a scale, or they can be *graphic indicators*. Likert-type scales, for example, provide a balanced list of alternatives from which raters may choose to indicate a response to a statement or question. The response chosen is attached to a quantity that becomes a score for use in data analysis. A rating item designed in this way might appear as shown below:

The instructor for this course was prepared for each class.

1. Strongly disagree
2. Disagree
3. Neither agree nor disagree
4. Agree
5. Strongly agree

This type of rating is so common that it appears intuitive, but it is not without its disadvantages. Concerns about of this method can arise when the data are analyzed statistically. Most researchers are comfortable treating these responses, and especially summed or averaged total scores across many such ratings, as *interval*-level measurement and apply parametric statistical methods, which assume that high level of measurement. Some critics point out that the psychological processes that lead to a rater's choosing a particular rating may not reflect an underlying interval-level scale (e.g., is the meaningful "distance" between *strongly agree* and *agree* ratings equal to the distance between *disagree* and *strongly disagree?*). So the use of parametric statistics may be inappropriate. The consensus of studies over the past several decades, however, suggests that the use of parametric statistical methods to analyze ratings of this type produces fairly accurate results, and the concern may be primarily academic.

Examples of graphic indicators as ratings include visual analog scales, which can be used to indicate with a high degree of precision a respondent's subjective impression. This method allows for a greater variability in responses and, theoretically, allows for the score to more closely estimate a true or accurate response. Pain scales, for example, ask patients to mark an "x" somewhere along a line to indicate the level of pain they are currently experiencing:

No pain————————x————Most pain I can imagine

The score for a graphic rating scale is determined by literally measuring the location of the mark along the line, in millimeters, for example, and that length becomes the rating. Beyond the potential for greater precision, an advantage of this rating method is that the measurement approach is more clearly interval- or ratio-level, and parametric statistical methods seem a clearly appropriate method of analysis. Concerns for this method center on the absence of many anchors, verbal description of scale points, and logistical measurement difficulties when forms are reproduced at different sizes, such as in different online contexts.

A third type of rating, the *checklist,* is sometimes quantified and sometimes represented as a set of independent observations. Checklists are lists of characteristics or components that may be present or absent in a person, object, or environment. A scale for depression, for example, or alcoholism, may ask people to place a checkmark next to, or circle, each characteristic that is true of them. The number of checks may be summed, and if a *cut score,* or critical total, is reached, then the raters may be categorized as, for example, being depressed or having a drinking problem.

The term *rating* is sometimes misused by researchers to describe a *ranking*. A rating is criterion referenced, which means the score has meaning when compared to the researcher-chosen definitions of different scale points; all elements or items may receive the same rating. A ranking is norm referenced: a score is defined as the rank order of the element or item when compared to other elements or items. With rankings, all elements, usually, must receive a different score. An advantage of ranking is that respondents may find it easier to rank order than to apply a rating to each element. A disadvantage of rankings is that less information is provided and, generally speaking, less powerful methods of statistical analysis are available for ordinal-level data. Researchers who wish to ultimately report rankings with their data can always measure their variables using ratings and determine the rankings if they wish at a later time, using the same data.

*Bruce Frey*

*See also* Levels of Measurement; Likert Scaling

**Further Readings**

Sax, G. (1996). *Principles of educational and psychological measurement and evaluation*. Belmont, CA: Wadsworth.
Thorkildsen, T. (2005). *Fundamentals of measurement in applied research*. Boston: Pearson/Allyn & Bacon.

# Ratio Scale

*Ratio scale* refers to the level of measurement in which the attributes composing variables are measured on specific numerical scores or values that have equal distances between attributes or points along the scale and are based on a "true zero" point. Among four levels of measurement, including nominal, ordinal, interval, and ratio scales, the ratio scale is the most precise.

Because attributes in a ratio scale have equal distances and a true zero point, statements about the ratio of attributes can be made. The score of zero in a ratio scale is not arbitrary, and it indicates the absence of whatever is being measured. Most variables in experimental sciences, particularly in behavioral sciences and natural sciences, are ratio-scale variables. One advantage of the ratio scale is that all mathematical operations are permitted for ratio-scale data.

## True Zero Point

Variables measured at the ratio-scale level have equal distances between attributes and a true zero point. The score of zero in ratio-scale variables indicates the absence or complete lack of an attribute. The variable *number of vehicles owned in the past 5 years* is an example of a ratio-scale variable. A score of zero for this variable means the respondent owned no vehicles in the past 5 years.

Similarly, it is possible for respondents to have zero years of work experience, no children, a score of zero on a midterm exam, zero dollars in a savings account, or zero people living below the poverty line. These are all examples of ratio-scale variables.

The true zero point also makes a ratio-scale variable more precise than an interval-scale variable. The true zero point allows ratio scale variables to make statements about the ratio of attributes. It is possible to multiply and divide the attributes of a ratio-scale variable. For example, it is possible to say that four vehicles are twice as many as two vehicles in the variable *number of vehicles owned in the past 5 years*.

Similarly, for the variable *number of years of work experience*, those who have 3 years of work experience have half as many years of work experience as those who have 6 years of work experience. For the variable *amount in a savings account*, $1,500 is 3 times as much savings as $500 or one third as much savings as $4,500. In the variable *proportion of people living below the poverty line*, areas that have 3,000 people living below the poverty line have twice as many people living below the poverty line as areas that have 1,500 people living below the poverty line.

## Mathematical Operations

One advantage, and perhaps the most important advantage, of the ratio scale is suitability for all mathematical operations. Data measurement on the ratio scale permits all mathematical operations to be used, including addition, subtraction, multiplication, division, and square roots. The use of all mathematical operations in turn permits the use of most statistical techniques for data measured on the ratio scale. There are many statistical techniques that are not appropriate for data measured at the nominal-, ordinal-, or interval-scale level but that are appropriate for the ratio-scale level, such as correlation, regression, factor analysis, and time-series analysis.

The statistics for ratio-scale variables are more powerful and produce more information than statistics for nominal-, ordinal-, and interval-scale variables. For example, statistics for measuring dispersion are available for nominal-scale data (index of qualitative variation), ordinal-scale data (range and the interquartile range), and interval- and ratio-scale data (standard deviation). Among the statistics for measuring dispersion, standard deviation is the most powerful statistic for measuring data dispersion.

## Dummy Variables

It is a common practice in statistics to convert nominal- or ordinal-scale variables to what are

called *dummy* variables so that they can be analyzed as if they were ratio-scale variables. In the simplest case, a dummy variable has two attributes, including 0 and 1. In experimental research design, a dummy variable is used to distinguish different treatment groups. A person is given a value of 0 if he or she is in the control group or a value of 1 if he or she is in the treatment group.

For example, the variable *gender* is a nominal-scale variable that has two attributes: female and male. Such a variable cannot be used in statistical techniques that are only appropriate for ratio-scale data unless the variable *gender* is converted into a dummy variable. In regression analysis, a nominal-scale variable such as *gender* can be used as an independent variable after being converted into a dummy variable. The attribute *female* could be coded as 1 and *male* could be coded as 0. After being converted into a dummy variable, the variable *gender* is no longer a nominal-scale variable and becomes the variable *femaleness*. The variable *femaleness* has a true zero point, and the attribute *male* has 0 *femaleness*. The variable *femaleness* could now be considered a ratio-scale variable and used as an independent variable in regression analysis.

*Deden Rukmana*

***See also*** Interval Scale; Nominal Scale; Ordinal Scale; Variable

**Further Readings**

Babbie, E. (2007). *The practice of social research* (11th ed.). Belmont, CA: Thomson Wadsworth.

Healey, J. F. (2005). *Statistics: A tool for social research* (7th ed.). Belmont, CA: Thomson Wadsworth.

Keyton, J. (2006). *Communication research: Asking questions, finding answers* (2nd ed.). Boston: McGraw-Hill.

Trochim, W. M. K. (2008). *The research method knowledge base: Dummy variable*. Retrieved December 14, 2008, from http://www.socialresearchmethods.net/kb/dummyvar.php

## Raw Scores

Raw scores are simply the total sum of a respondent's selections before any statistical conversions are made. While serving as rudimentary level of statistical measurement, raw scores are functionally limited. In application, raw scores are used to arrive at a set of standardized scores (i.e., *T* scores, *z* scores) that can be used to compare individuals to a reference group. Researchers use raw scores to perform statistical analyses or to norm measures. Applied practitioners use raw scores to communicate performance or measurement results.

Although not universally true, raw scores typically are the sum of correct responses out of the total possible correct responses. For example, on a scale containing 10 questions, a respondent may correctly answer 8. Therefore the respondent would achieve a raw score of 8. Of course, this raw score converts rather easily to the percentage correct on the scale, in this case 80%. In this example, 80% is a representation of the respondent's raw score; however, only the number 8 is considered the actual raw score.

Educational measurement is a common application of raw score usage. Take for instance a student who achieves a raw score of 17 on an assessment. This score provides limited information without an indicator of the total possible score. If the total score was 20, one would have a better indicator of the student's performance, yet this assessment provides no information about the student's performance relative to peers (i.e., normative) or to the student's past performance (i.e., growth modeling). However, if these other scores are part of the student's ongoing classroom assessment, one would have enough information to tally the student's current performance level. Even if this particular assessment weighs more or less heavily than other assessments with a total of only 20 possible points, these scores provide enough information to allow us to create a total of points that go into the student's overall performance file.

Applications of raw scores do not always require a measure with correct or incorrect responses. Measures of interest, personality, or motivation, for example, do not contain right or wrong answers, but responses that reflect the participant's response preferences. In this example, these inventories measure a number of domains, usually through a Likert-type scale; scores on these domains are summed but do not represent a ratio of correct to incorrect responses.

Although in most cases raw scores provide enough information for typical educational purposes, they remain limited in measurement applications. Consider for a moment that we want to examine how this student performed in comparison to peers. In a class with a small number of students, it might be tempting to simply order the performance of peers from lowest to highest and perform a count of the scores until we arrive at our student's score. However, if the scores were taken from a class of five and those additional four scores were 19, 16, 14, and 17, simply ordering those scores and then counting until we arrive at our student's score becomes inadequate. If one is interested in the student's score in comparison with peers, one would consider the group's highest and lowest score on this assessment and remove the full range of possible scores from consideration. Through the use of a selected statistical metric, we are able to convert the student's raw score into a standard score and determine how it compares to the scores of peers.

As another example, imagine a student who has taken a standardized computer adaptive assessment as part of graduate school application requirements. The test has been constructed and normed by the use of stringent testing standards. Under computer adaptive testing conditions, test item difficulty for the subsequent question depends on the student's previous performance. As the student answers questions correctly, the difficulty of test items increases. Conversely, when the student misses a question, the testing program responds by selecting an easier item. This type of testing design allows efficient administration while obtaining a valid measure of the student's ability. However, for two students who both responded to 60 questions of varying difficulty, it becomes impossible to use raw scores alone to determine how the students rank in comparison with each other. Assume that these students both correctly respond to 45 questions out of 60, but due to differences in the two students' abilities, Student A correctly answers the first few questions before missing an item, whereas Student B incorrectly answers the initial questions. In this situation, Student A will continue to face increasingly difficult questions while Student B receives increasingly easier questions. Although both students correctly answer 45 questions out of 60, Student A correctly answers more questions at a higher difficulty level. Because of the differing levels of item difficulty, the raw scores of correctly answered questions does not provide an accurate reflection of the two students' performance.

One could argue that once those scores are correctly weighted, the raw scores resulting from those exams provide enough information to determine which student performed better. However, this assumption applies only in tests in which each item contains clear weighting. Let us now imagine that Student A has applied to a competitive program and is now competing for admission with other students who took tests from the same testing company, measuring the same domain; however, the test is frequently reconstructed to maintain test security, and each test reconstruction is an essentially different test. Because these tests used different questions in their item pools, it becomes impossible using raw scores alone to determine how those scores compare with each other. In this example, Student A is now competing with students who achieved scores ranging from 15 points higher to 15 points lower than A's score of 45. Even assuming the total possible correct score for all students is 60, information on which to weight the tests is still lacking. Comparing student scores more accurately requires further statistical information, including the mean and the standard deviation scores. As previously mentioned, Student A has correctly responded to 45 questions and is now competing for admission with Student C, who has obtained 50 out of 60 points on the exam. For convenience, let us assume that all items are weighted equally. Based on these scores alone, it is tempting to assume Student A performed inferiorly compared with Student C. However, on further examination, we learn that the mean score (i.e., arithmetic average) from the test our student took was 40, and the standard deviation score is 5 points. With this information we now understand that Student A scored 1 standard deviation above the mean and higher than or equal to approximately 84% of the students who took the same exam. An analysis of Student C's scores reveals the mean score from Student C's test is 55 and the standard deviation score is 2 points. Now it becomes clear that even though Student C has a raw score that is 5 points higher than Student A's score, Student C actually scored 2.5

standard deviations below the mean. Assuming there were no biases in the assessment, we can conclude that even though Student A has a lower raw score, Student A is likely a significantly better candidate than Student C in this particular domain.

*Justin P. Allen*

***See also*** Mean; Standard Deviation; Standardized Score; *z* Score

**Further Readings**

Siegle , D. (n.d.). *Standardized scores*. Retrieved March 7, 2010, from http://www.gifted.uconn.edu/siegle/research/Normal/Interpret%20Raw%20Scores.html

## Reactive Arrangements

Reactive arrangements are an example of a threat to the internal validity of a research design. *Reactivity* occurs when the results of an experiment are due, at least in part, to behaviors of participants that artificially result from their participation in the experiment. Thomas D. Cook and Donald T. Campbell and colleagues, in their works defining threats to validity in research design, distinguish between artifacts due to the use of measures in a study and other reactions that occur because participants are aware that they are in a study. When these reactions become a functional part of the treatment or independent variable, then reactive arrangements are present.

Reactivity has occurred when the "meaning" of a treatment includes the human reactions to being in a study. Reactions to the study procedures themselves may occur in several different ways:

- Participants may respond favorably after receiving a nonactive drug used as a control (the placebo effect).
- Study subjects may guess or ascertain the expected outcome of an experiment and behave in ways they feel will please the researcher.
- Participants may perform higher or lower on skill or achievement measures because of increased motivation or anxiety-induced interference.
- Apprehension of being evaluated by a "scientist" may encourage some subjects to produce expected responses.

Protections against the threat of reactive arrangements center on hiding the true purpose of a study by making all control treatments appear to be authentic, measuring the outcome surreptitiously, and designing pretest measures so as not to give cues for expected outcomes. A researcher examining the effectiveness of a psychological therapeutic intervention, for example, may replace a no-treatment group with a group that receives another approach to therapy that is believed to be weaker or ineffective. Many studies use this approach with their treatment-as-usual group. Another approach is for researchers to make the measurement of outcomes less obvious by not assessing the outcome at the immediate conclusion of the treatment or procedures but, if possible and still theoretically meaningful, assessing outcomes on some delayed basis.

Participants will almost always make their own hypotheses as to the purpose of a study, however, so complete protection against this threat may not be possible. There are also ethical requirements to give enough information about the purpose of a proposed study so that recruits can give free and informed consent to take part. Some extreme ethical positions might even require that the researcher share the study hypotheses with all recruits. Most researchers, however, control for the threat of reactive arrangements through design choices or, at the least, measure variables in ways that allow for a determination of the presence of reactive arrangements.

Reactive arrangements are a particularly tough threat to compensate for, because they are among those threats to validity that are not eliminated or lessened by random assignment to different groups. Random assignment can control for preexisting differences between members of different groups, but in the case of reactive arrangements, the differences between groups do not exist until the study begins. While group comparison is the basis of modern experimental research, one consequence of this analytic technique is the possibility of reactive arrangements.

*Bruce Frey*

*See also* Ethics in the Research Process; Placebo Effect

**Further Readings**

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Boston: Houghton Mifflin.
Shedish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

# RECRUITMENT

The success of any human subject research project is usually only as good as the researcher's ability to recruit the proper participants. Common limitations of empirical research studies usually include small sample sizes and/or samples that are not representative of the population. Small and/or nonrepresentative samples could result for a number of reasons (e.g., limited funding, time, access to the population of interest); however, effective recruitment is often a factor. Recruitment is the act of enlisting people for a certain cause, in this case, participation in a research experiment. This entry discusses different methods of recruitment and strategies to improve recruitment methods.

## Recruitment Methods

Considerations and planning for recruitment should occur during the early stages of project development. A methodology section often needs to be tailored to accommodate the needs of recruitment and retention (keeping individuals in the study). The ability to attract and maintain participants is crucial to most studies, and some methodologies can lead to problems with participation and recruitment. Methodological problems may include length of time of the study, number of surveys, location, and language and literacy barriers. In addition, partnerships formed with community agencies, which will increase recruitment, may have restrictions that require alterations to the methods. For this reason in particular, designing recruitment strategies needs to occur at the beginning of the study design to allow for any necessary adjustments.

There are several ways to recruit participants for research studies. One common method used at universities is department subject pools. In several undergraduate classes, students are often mandated to participate in research studies for course credit, and students are provided a list of research projects available for participation.

Online research advertisements are also very popular. There are several options for posting online research. For example, popular social networking sites such as MySpace or Facebook are often used. Researchers can also contact website owners of relevant topics and ask for permission to advertise. Depending on the site, an advertising fee may be required. There are also several websites dedicated to posting research studies. If a study is posted online but requires in-person contact, it is advisable to make sure that the advertisements are located on a website frequented by individuals within the geographical area.

Other common recruitment strategies include advertisements in newspapers, community magazines, television or radio announcements, flyers posted at community centers, stores, related businesses, or available at related community events. Some of these forms of advertisement may be costly; however, companies may be willing to donate space, especially if the researcher is affiliated with a nonprofit organization such as a public university. It is important that before materials are posted in public places, permission for posting is obtained.

Another form of recruitment includes the use of a preexisting database of individuals who have consented to be contacted to participate in research studies. University departments or community agencies can be contacted to inquire about requirements to access these databases. Examples include a registry established at the Center for Cognitive Neuroscience at the University of Pennsylvania and the Cognitive Neuroscience Research Registry at McGill University, which provide a structure for recruitment and retention and serve as a data management system for cognitive neuroscience researchers. However, as with these examples, some of the participant lists may be specific to one topic.

Many studies use multiple methods of recruitment to ensure that the study consists of the target population and the desired sample size. However,

it is not necessary to use all forms of recruitment in every study to be successful. There are limitations, such as time and funding, that dictate whether some recruitment methods are feasible for a study. What is most important is that the recruitment methods used are tailored specifically to the target population in order to maximize enrollment in the research study.

## Strategies for Improving Recruitment

There are several strategies that researchers can employ to encourage participation in a study. Each of these strategies can be used alone or in conjunction with one another, depending on the circumstances of the researcher.

### *Tailoring Recruitment to the Population and Type of Study*

Two key factors of a study that have a major influence on recruitment strategies are the target population (e.g., children, older adults, ethnic minorities, abused women) and the specific methods involved in the study (e.g., survey instruments, treatment, longitudinal design). Recruitment has been known to be challenging for certain target populations, one being minority populations. Common reasons for lack of participation often include inappropriate methodology used to reach the target community; recruitment materials that are not understood by the target population; a perception by the target population that the research is not relevant; lack of community involvement by researchers, leading to lack of community trust; no minority representation on research teams; and few incentives provided for participation. Issues such as these must be addressed and their solutions built into the research design to ensure ample recruitment. With regard to recruitment challenges due to unavoidable research methodology, such as longitudinal studies or physically invasive medical research, challenges could be due to higher demands that are placed on the participants.

Judith McFarlane in 2007 conducted a longitudinal study on abused women and found that retention rates were between 89% and 100% when efforts were made to establish a strong working relationship with the community agencies, incremental monetary incentives were offered, contact was made with the participants on a constant basis, and safety factors were taken into consideration. This study had two complications: working with a vulnerable population and the requirement of a long-term commitment. Each research study may require special accommodations to entice and retain participants. For this particular study, it was crucial to ensure safety for participants, and team members had to have consistent personal contact in an effort to decrease drop-out rates.

### *Community and/or Organizational Involvement*

One barrier to recruitment is often mistrust, especially in vulnerable populations (children, minorities, abused women, etc.). If a partnership is established with a respected community member, agency, and/or organization, individuals within that community may be more likely to participate. Recruitment can also be easier for the researcher when a connection is formed with a group that is willing to perform some of the recruitment activities. Some common community connections could be through community leaders, local agencies or community centers, churches, or local schools. One example is forming a relationship with a target community school whose principal would endorse the research and encourage staff to pass out flyers, speak to parents, and so on. In this example, the recruitment occurs through a source in which the individuals may already have an established relationship.

Collaboration with an agency (a company or an organization) for recruitment purposes can have a significant impact on the success of the study. It is important to learn about the agency and the community the agency serves in order to build a good working relationship and determine goodness of fit for the research study. Building this relationship may include spending a day at the agency and speaking with staff and current clients. It is important to keep the agency informed and to offer something to the agency in return for its assistance. This can be a monetary payment or an exchange of services, depending on the agency's need and the resources of the researcher. For example, researchers can offer to teach classes or assist with problems in the researchers' area of expertise. One potential problem with agency

collaboration is the possibility that agency policies and procedures could affect the research study. Offering to have a member of the agency on the research team may allow the agency to feel more invested in the project and to maintain consistent contact regarding progress or problems with the research study, and it may result in fewer problems when determining adjustments to the research design to comply with agency rules. If having a member of the agency on the research team is not possible, then a member of the research team must connect with the agency on a regular basis. Oftentimes in research, several research team members and assistants may need to interact with agency workers, which can cause confusion and frustration for the agency. It may be beneficial to have as assigned agency contacts a small number of researchers who are willing to work with the agency's schedule.

### *Personalized Recruitment*

Recruitment methods that are personal elicit better responses than impersonal methods do. Strategies include having research assistants phone individuals, walk door to door in target neighborhoods, set up information tables at community events or centers, or speak to individuals at target community centers, shopping malls, or agencies. It is always important to obtain permission from the property owner if advertising will be conducted on private property. Another important task in personalized recruitment is accurate tracking logs that list the date, time, person making the contact, and person contacted. This will avoid a single participant's receiving multiple telephone calls or visits, which may discourage participation. Even though these methods may be time-consuming, personalized strategies may increase trust, acceptance, and the likelihood that certain populations are aware of the research study.

### *Language*

Language can serve as a barrier for many different populations for different reasons. Some populations may not speak the language used in the recruitment strategy, the language may be too complex to fully explain the study, the print may not be legible, or the advertisement design may be too distracting. Overall, terminology should always be as simple as possible, and researchers may want to avoid the use of certain words that might offend potential participants. For example, advertisements should have a title using general, familiar terminology in addition to a more detailed description of the actual study. The use of nonthreatening words and words that carry no stigmas may make potential participants more comfortable. It is important that these adjustments do not deceive participants but rather provide a clear explanation of the study. This same idea can be used when introducing consent forms. In some studies, consent forms are sent home as part of the recruitment process. These forms can be very formal, hard to understand, and intimidating. Even though a consent form may not allow for alteration of language, a cover letter can often be added that is written for the target audience and can explain the technicalities in simpler terms.

### *Transportation and Child Care*

In many families, participation is difficult because of lack of transportation or child care. Using available funding to provide these resources or conducting the study at a convenient time for the participant or at the participant's home, when possible, could significantly increase participation in certain populations.

### *Incentives*

Providing compensation can be a great recruitment method. It is the ethical responsibility of the researcher to ensure that the compensation is not coercive. Incentives can come in the form of gift cards, cash, or other materials (i.e., food, clothes). Researchers could use local or national wage standards to determine dollars per hour for participation. It has been previously suggested that monetary incentives used for longitudinal studies be provided in an incremental way, whereby each time a participant completes a part, the amount increases. It is very important to pick incentives that match the goals or needs of the population; for example, if a study is being conducted on the homeless population, it may be appropriate to provide food or clothing. Including incentives on

advertisement materials can enhance recruitment efforts.

### Retaining Participants

An effective recruitment plan will be responsive to the needs of the target population and the specifics of the research design. If the above suggestions are taken into consideration as the methods are designed, recruitment should prove to be less of a challenge and a realistic understanding of the ability to obtain certain samples will be reached in advance. The recruitment process does not end once an individual has agreed to participate. Recruitment strategies need to be maintained throughout the study in order to retain participants.

*Amanda Haboush*

***See also*** Ethics in the Research Process; Experimental Design; Longitudinal Design; MBESS; Methods Section; Participants; Planning Research; Research Design Principles; Sample; Sample Size; Volunteer Bias

#### Further Readings

Cauce, A. M., Ryan, K. D., & Grove K. (1998). Children and adolescents of color, where are you? Participation, selection, recruitment, and retention in developmental research. In V. V. McLoyd & L. Steinberg (Eds.), *Studying minority adolescents: Conceptual, methodological, and theoretical issues* (pp. 147–166). Mahwah, NJ: Lawrence Erlbaum.

Kirchhoff, K. T., & Kehl, K. A. (2008). Recruiting participants in end-of-life research. *American Journal of Hospice Palliative Care, 24,* 515–521.

McFarlane, J. (2007). Strategies for successful recruitment and retention of abused women for longitudinal studies. *Issues in Mental Health Nursing, 28,* 883–897.

Prinz, R. J., Smith, E. P., Dumas, J. E., Laughlin, J. E., White, D. W., & Barron, R. (2001). Recruitment and retention of participants in prevention trials involving family-based interventions. *American Journal of Preventive Medicine, 20*(Suppl. 1), 31–37.

Rubin, A., & Babbie, E. R. (2008). *Research methods for social work* (6th ed.). Belmont, CA: Thomson Brooks/Cole.

Sullivan, C. M., Rumptz, M. H., Campbell, R., Eby, K. K., & Davidson, W. S. (1996). Retaining participants in longitudinal community research: A comprehensive protocol. *Journal of Applied Behavioral Science, 32,* 262–276.

## Regression Artifacts

The term *regression artifacts* refers to pseudoeffects from a regression type of analysis. These incorrect causal estimates are due to biases from causes other than the cause of interest. Note that such artifacts are problems only when making causal inferences, not when merely predicting some future outcome. A famous regression artifact, for example, was that the first major evaluation of Head Start concluded that the summer program actually caused children to do worse in school. If this were merely a prediction, it would have been correct that having attended Head Start would predict poorer performance in school. These children did poorly in school because of their disadvantaged status, which Head Start could not completely compensate for. This correct prediction becomes an artifact only when someone makes a *causal* conclusion that Head Start is responsible for their below-average academic performance later.

The most general reason for regression artifacts is that the statistical analysis reflects an incomplete picture of reality, which is called a *specification error.* Specification errors include the omission of relevant variables and other mismatches between statistical assumptions and reality. A *relevant variable* is any variable that is associated with the cause of interest but that also causally influences the outcome variable directly. In the Head Start case, being from a disadvantaged background was associated with attendance at Head Start, but it also caused poor academic performance. Mismatches between reality and linear statistical assumptions include a curvilinear relationship and an interaction effect, in which the effect of one causal variable depends on another variable. Another common mismatch is that unbiased causal evidence usually requires that other relevant variables (potential confounds) be measured without error. Socioeconomic status (SES) was controlled for statistically in the first major evaluation of Head Start, but measurement error in SES reduced its ability to fully correct for disadvantage. Regression estimates of causal influences are unbiased only if statistical analyses include perfectly valid and reliable measures of all relevant variables and correctly reflect any complexities among their

interrelationships. To the extent a regression type of analysis falls short of that ideal, there is potential for regression artifacts, that is, pseudoeffects masquerading as causal evidence.

## Types of Regression Artifacts

### *Regression Toward the Mean*

There are several major types of regression artifacts. The most common is regression toward the mean, which is summarized in another entry in this Encyclopedia, as well as in an important book by Donald Campbell and David Kenny, titled *A Primer on Regression Artifacts*. Regression toward the mean occurs when the participants are selected on the basis of their extreme scores (e.g., deciding to start therapy because of a bad day). Such people usually move toward the mean spontaneously, which could be incorrectly interpreted as a causal effect, such as improvement due to therapy.

### *Underadjustment Bias*

Another example of regression artifacts is underadjustment bias, the systematic error remaining after typical statistical adjustments for potential confounds. Campbell and Kenny have shown that several kinds of statistical adjustments reduce the bias but do not eliminate it. They called this one of the most serious difficulties in data analysis. In the Head Start example, statistical adjustments for differences in SES gave the illusion that the effects of disadvantage were removed, even though they were only reduced. This is why Head Start appeared to be hindering students' subsequent success in school, even when that was not the case. It was the residual (remaining) influence of being disadvantaged, and not Head Start, that was hindering students' success in school.

### *Attenuation*

A third type of artifact is attenuation, which is the reduced magnitude of regression coefficients due to measurement error or restricted range in the variables. Attenuation can be corrected by dividing the correlation by the square root of the product of the two variables' reliabilities. The accuracy of the correction requires the correct estimate of reliability. When regression analyses include multiple predictor variables, measurement error and restriction of range in one predictor can cause inflation as well as attenuation in coefficients for other predictor variables (e.g., the underadjustment bias).

### *Suppression Effect*

The final example of a regression artifact is a suppression effect. The main concern for most other regression artifacts is that a significant association might falsely suggest a causal effect. In contrast, a suppression effect occurs when a causal variable appears to be unrelated to an outcome, whereas in fact, its causal effect has been suppressed by other causal influences that are not represented in the regression analysis. A classic example is that paper-and-pencil aptitude tests might fail to predict a practical ability, such as potential ability to be a competent airplane pilot. In this hypothetical example, doing well on the aptitude test might reflect verbal ability as well as potential piloting ability. In a suppression effect, adjusting statistically for verbal ability would then show the aptitude test to predict future piloting ability, even though there was no association between those two variables before the statistical adjustment.

## Adjustments

Overall, regression artifacts are possible from regression types of analyses unless groups have been equated successfully on all characteristics except the causal variable being investigated. Random assignment is the best way to equate the groups. When that cannot be done, there are several ways to adjust statistically for potential confounds or causes, but most adjustments reduce the bias without eliminating it. Therefore, studies that rely solely on statistical adjustments need to consider other plausible explanations of their results in order to alert themselves and their readers to the possibility that their causal evidence might be misleading due to a regression artifact.

*Robert E. Larzelere and Ketevan Danelia*

*See also* Confounding; Correction for Attenuation; Multiple Regression; Regression to the Mean; Restriction of Range; Statistical Control

**Further Readings**

Campbell, D. T. (1996). Regression artifacts in time-series and longitudinal data. *Evaluation & Program Planning, 19,* 377–389.

Campbell, D. T., & Kenny, D. A. (1999). *A primer on regression artifacts*. New York: Guilford Press.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

# Regression Coefficient

The regression coefficient expresses the functional relationship among the response (explained, dependent) variable and one or more explanatory (predictor, independent) variables. Denoting the response variable by $Y$ and the set of explanatory variables by $X_1, X_2, \ldots, X_k$, the regression model can generally be formulated as

$$Y = f(X_1, X_2, \ldots, X_k) + \varepsilon,$$

where $k$ denotes the number of predictor variables and $\varepsilon$ denotes the random disturbance or error, representing the discrepancy between the observed response variable and the estimated regression line.

Following the commonly used notational convention in linear regression analysis, which uses Greek letters to denote the unknown parameters, the linear regression model can be written as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon,$$

where $\beta_0$ denotes the intercept and $\beta_1, \beta_2, \ldots, \beta_k$ denote the regression coefficients to be estimated from the data. More specifically, the regression coefficients indicate the dependence of the response variable on one or more explanatory variables, as shown in the linear regression model above. The parameter $\beta_0$ denotes the intercept, or where the regression line crosses the $y$-axis. As such, the intercept $\beta_0$ determines the mean value of the response variable $Y$, independent of any explanatory variable.

## Estimating Regression Coefficients

### *Purposes of Estimating Regression Coefficients*

The estimated regression coefficients or "estimators" $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ (the "hat" on the beta, i.e., '^', refers to an estimated regression coefficient), calculated from the regression analysis, can then serve multiple purposes:

1. *Evaluating the importance of individual predictor variables.* A common approach here would be to use standardized response and predictor variables in the regression analysis. From the standardized variables, one can estimate standardized regression coefficients, which measure in standard deviations the change in the response variable that follows a standard unit change in a predictor variable. The standardization procedure allows one to assess the importance of individual predictor variables in situations in which the predictor variables are measured in different units.

2. *Analyzing the changes in the response variable following a one-unit change in the corresponding explanatory variables $X_1, X_2, \ldots, X_k$.* For instance, $\hat{\beta}_1$ indicates the change in the response variable following a one-unit change in the predictor variable $X_1$, while holding all other predictor variables, $X_2, X_3, \ldots, X_k$, constant.

3. *Forecasting values of the response for selected sets of predictor variables.* In this case, using all the estimated regression coefficients, $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$, together with a given set of predictor variables $X_1, X_2, \ldots, X_k$, one would be able to forecast the expected value of the response variable.

Besides the absolute magnitudes of the estimated regression coefficients, the signs of the estimators are of major importance because they indicate direct or inverse relationships between the response variable and the corresponding predictor variables. It is usual to have prior expectations of the expected outcomes of the signs of the estimators. In other words, it is conceptually predetermined whether the relationship between the response variable and a predictor variable should be of positive or negative nature.

### *Example of Use of Estimated Regression Coefficients*

In real estate appraisal, an analyst might be interested in how building characteristics such as floor space in square feet ($X_1$), number of bedrooms ($X_2$), and the number of bathrooms ($X_3$) help in explaining the average market values ($Y$) of real estate properties. In this oversimplified example, the estimated regression coefficients may be used for several purposes. Using standardized regression coefficients, one could determine whether the floor space in square feet, the number of bedrooms, or the number of bathrooms is most important in determining market values of real estate properties. In addition, the unstandardized, estimated regression coefficient $\hat{\beta}_1$ would indicate the expected change in the average market value of real estate properties for a 1-square-foot change in floor space while the number of bedrooms and bathrooms remain constant. Analogously, $\hat{\beta}_2$ or $\hat{\beta}_3$ would indicate how much one more bedroom or one bathroom would potentially add to the expected market value of real estate properties. And finally, using $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$, one could estimate the market value of a hypothetical 2,200-square-foot real estate property with three bedrooms and two bathrooms. Of course, one would expect all the estimated regression coefficients to have positive signs.

### *Method of Estimating Regression Coefficients*

The regression coefficients $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ are estimated from the available data. As a general rule, the number of observations ($n$) must be greater than the number of explanatory variables or regression coefficients ($k$) to be estimated, and the more the better. Note that the estimation procedure is stochastic—rather than deterministic—by nature. Thus, the inclusion of the random disturbance term is essential to account for the discrepancy between the observed response variable and its corresponding estimated value. Consequently, the assumptions embedded in the disturbance terms imply important and far-reaching properties to the estimated regression coefficients.

The most commonly used and conceptually most easily understood method of estimating the regression coefficients is the ordinary least squares (OLS) method. Accordingly, the estimated regression coefficients are often referred to as OLS regression coefficients, or $\hat{\beta}^{\mathrm{OLS}}$. The least squares regression coefficients are obtained by minimizing the sum of squares of vertical distances from each observation to the fitted line. These vertical distances, the differences between the actual and estimated $Y$ values, or $Y_i - \hat{Y}_i$, represent the errors in the response variables, or the variation in the response variable that cannot be explained by the regression model.

Using the simple, one-predictor-variable regression model,

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon,$$

and applying the process of differential calculus to minimize the squared vertical distances yields the following normal equations:

$$\sum Y_i = n \cdot \hat{\beta}_0 + \hat{\beta}_1 \cdot \sum X_i$$

$$\sum Y_i X_i = \hat{\beta}_0 \sum X_i + \hat{\beta}_1 \cdot \sum X_i^2$$

Solving these two normal equations simultaneously, we get the solutions for the two OLS regression coefficients for the simple regression model:

$$\hat{\beta}_1 = \frac{\sum (X_i - \overline{X}) \cdot (Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2} = \frac{\sum x_i \cdot y_i}{\sum x_i^2}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X},$$

where $n$ denotes the total number of observations, and $i$ denotes an individual observation between 1 and $n$.

## Properties of Estimated Regression Coefficients

### *Gauss–Markov Theorem*

The properties of estimated regression coefficients, in return, depend on the assumptions embedded in the disturbance terms. The *Gauss–Markov theorem* states the following assumptions

for the disturbance terms in a linear regression model: (a) the mean value of disturbance terms is zero, or $E(\varepsilon_i) = 0$; (b) all disturbances have the same variance, or $\text{Var}(\varepsilon_i) = \sigma^2$, which is also referred to as homoscedasticity; and (c) all disturbance terms are uncorrelated with one another, or $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$, for $i \neq j$. Following the Gauss–Markov assumptions of the disturbance terms, the estimated OLS regression coefficients for the simple regression model have the following properties:

1. The estimated OLS regression coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased:

$$E(\hat{\beta}_0) = \beta_0$$

$$E(\hat{\beta}_1) = \beta_1.$$

2. The estimated OLS regression coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ have minimum variance:

$$\sigma^2_{\hat{\beta}_0} = \frac{\sum X_i^2}{n \cdot \sum x_i^2} \cdot \sigma^2$$

$$\sigma^2_{\hat{\beta}_1} = \frac{\sigma^2}{\sum x_i^2}.$$

Putting it together, the Gauss–Markov theorem implies that for a linear regression model—in which the disturbance terms are uncorrelated and have a mean value of zero and equal variances—the unbiased OLS estimators of $\beta$ have minimum variance. In this case, the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are said to be the *best linear unbiased estimators* of $\beta_0$ and $\beta_1$.

### *Disturbance*

So far, no assumptions have been made regarding the distribution of the disturbance terms. Assuming that, in addition to the Gauss–Markov theorem, the disturbance terms are normally and independently distributed (NID), or

$$\varepsilon_i \sim \text{NID}(0, \sigma^2),$$

adds the following properties of estimated regression coefficients to previous results:

1. The estimator $\hat{\beta}_0$ is normally distributed:

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2_{\hat{\beta}_0}\right).$$

2. The estimator $\hat{\beta}_1$ is normally distributed:

$$\hat{\beta}_1 \sim N\left(\beta_1, \sigma^2_{\hat{\beta}_1}\right).$$

3. The estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are distributed independently of $\hat{\sigma}^2$.
4. The estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are consistent estimators. With increasing sample size $n$, $\hat{\beta}_0$ and $\hat{\beta}_1$ converge to their true, but unknown, population parameters.
5. The quantity $\frac{(n-2)\cdot\hat{\sigma}^2}{\sigma^2}$ has a chi-square distribution with $(n - 2)$ degrees of freedom.
6. $\hat{\beta}_0$ and $\hat{\beta}_1$ are now the *best unbiased estimators* of $\beta_0$ and $\beta_1$. Relaxing the linearity assumption of the estimators, they have now minimum variance of all unbiased estimators.

### *Testing Estimated Regression Coefficients*

The normality assumption of the random disturbance terms implies important and far-reaching properties to the OLS regression estimators. The most important is that $\hat{\beta}_0$ and $\hat{\beta}_1$ follow a normal probability distribution, and $\hat{\sigma}_2$ follows a chi-square distribution. The same regression coefficient properties, of course, hold for the multiple linear regression model.

With their sampling distributions, the estimated regression coefficients can then be tested for their statistical significance. For a predictor variable $X_1, X_2, \ldots, X_k$ to be a statistically significant predictor of the response variable $Y$, it must be different from zero. For the simple regression model, the null hypotheses for testing the usefulness of the two estimators can be written as

$$H_0 : \beta_0 = 0 \text{ and}$$

$$H_0 : \beta_1 = 0.$$

The alternative hypotheses would then be $H_1 : \beta_0 \neq 0$ and $H_2 : \beta_1 \neq 0$. Given that in most cases the true population variance $\sigma^2$ is not known and is therefore replaced in practice by the unbiased estimator $\hat{\sigma}^2$, the Student's $t$ distribution is the

appropriate test distribution with $(n - 2)$ degrees of freedom for the simple regression model. The test statistics are computed as

$$t_0 = \frac{\hat{\beta}_0}{se(\hat{\beta}_0)}$$

and

$$t_1 = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)},$$

where $se(\hat{\beta}_0)$ and $se(\hat{\beta}_1)$ denote the standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively. The standard error is a measure on how precisely the regression coefficients have been estimated; the smaller the standard error, the better. The standard errors for the two estimators of $\hat{\beta}_0$ and $\hat{\beta}_1$ are calculated as

$$se(\hat{\beta}_0) = \hat{\sigma} \cdot \sqrt{\frac{\sum X_i^2}{n \cdot \sum x_i^2}}$$

$$se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum x_i^2}}.$$

The hypothesis test for each estimator is carried out by comparing the calculated test statistics $t_0$ and $t_1$ with the appropriate critical value from the $t$ table in a two-tailed test. The null hypothesis, which states that the estimated regression coefficients are zero, can be rejected at a selected level of significance of $\alpha$ if the test statistic falls in the critical rejection region in either of the two tails of the test distribution; that is, if the absolute value of the test statistic is larger than or equal to the appropriate critical value. For the estimator $\hat{\beta}_1$, for instance, the decision criterion is

$$|t_1| \geq t_{(n-2,\,\alpha/2).}$$

Using the $p$-value approach, the criterion, which is equivalent to the $t$-test criterion above, to reject $H_0$ can be formulated as

$$p(|t_1|) \leq \alpha,$$

where $p$ denotes the area left in the two tails that corresponds to the calculated test statistics.

Regression software packages provide tests of significance for each estimator as standard output. Usually, each estimator is reported with its corresponding $t$ statistic and standard error. In addition, an $F$ test, which tests the overall statistical significance of the regression model, is usually provided. The null hypothesis for the $F$ test states that all predictor variables (excluding the intercept) under consideration are simultaneously zero; in other words, they have no explanatory power. This is stated as

$$H_0 : \beta_1 + \beta_2 + \cdots + \beta_k = 0.$$

The alternative hypothesis $H_1$ for this specific case then states that not all regression coefficients are simultaneously zero. Following the same logic as for the $t$ test, $H_0$ can be rejected if the calculated $F$ statistic is larger than the critical $F$ value for the given level of significance, where the $F$ statistic is calculated as

$$F = \frac{\text{explained sum of squares}/k}{\text{unexplained sum of squares}/(n-k-1)},$$

where the explained sum of squares is due to regression and the unexplained sum of squares is due to error. The two sums of squares needed for the $F$ test are reported in the analysis of variance table.

In practice, the most widely used approach to estimate the regression coefficients is the OLS approach. Alternative methods include the Bayesian method, the maximum likelihood approach, quantile regression, nonparametric regression, and nonlinear regression.

*Rainer vom Hofe*

***See also*** Logistic Regression; Multiple Regression; Regression Artifacts; Stepwise Regression

**Further Readings**

Chatterjee, S., & Hadi, A. S. (2006). *Regression analysis by example* (4th ed.). Hoboken, NJ: Wiley.

Greene, W. (2008). *Econometric analysis* (6th ed.). Upper Saddle River, NJ: Prentice Hall.

Gujarati, D. N. (2002). *Basic econometrics* (4th ed.). New York: McGraw-Hill.

Kennedy, P. (2008). *A guide to econometrics* (6th ed.). Hoboken, NJ: Wiley-Blackwell.
Tabachnik, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston: Allyn & Bacon.

# Regression Discontinuity

Regression discontinuity (RD) is a two-group, pre-post, quasi-experimental method for evaluating the effect of an intervention. Assignment to the experimental or comparison group is based on the value of some variable. That variable can be either the baseline value of the outcome itself or some other variable. The effect of the intervention is assessed by looking at the discontinuity of the regression lines between the groups at the cutoff point.

## History

The RD design was introduced in 1960 by Donald Thistlewaite and Donald Campbell to evaluate the effects of compensatory educational interventions. Since that time, it has been used in areas such as criminal justice, social welfare, economics, and pharmaceutical health services research, although it is still a relatively rarely used design. Indeed, a recent search of the psychological and medical literature since 1960 turned up only 59 references, of which 34 were about the technique and only 25 actually used it. Of the latter, about one third were dissertations, and nearly three quarters were in the field of education.

## Method

The unit of assignment can be individuals or groups, such as classes within a school, entire schools, or even school districts. The units are assigned to the treatment or control conditions on the basis of their score on some measure. The cutoff score can be determined in one of two ways: either on the basis of knowledge of the properties of the score (e.g., a failing grade, the proportion of high school students not graduating, or blood pressure above some threshold for hypertension) or on the basis of the resources that are available (e.g., the program is able to handle only 25% of the people). In the simplest case, the grouping variable is the outcome variable itself, which is measured before and after the intervention for the experimental group and at equivalent times for the comparison group. Hence, the basic design looks like

| | | | |
|---|---|---|---|
| *Experimental group* : | Pretest | Intervention | Posttest |
| *Comparison group* : | Pretest | | Posttest |

and the results are plotted with the pretest on the x-axis and the posttest on the y-axis.

For example, people can be chosen for a weight-loss program if their body mass index (BMI) is more than 30, which is regarded as the cut point for obesity. At the end of a 6-month intervention, the BMIs for all people are measured again, and the hypothetical results are shown in Figure 1.

If the intervention has no effect, then the regression lines will be continuous at the cut point. The discontinuity at the cutoff, seen in Figure 1, indicates that the program was effective in that the postintervention scores for the experimental group tend to be lower than the prescores, whereas the before and after BMIs for those in the comparison group are about the same.

It is not necessary that the pre- and posttreatment variables be the same or that the cut point be

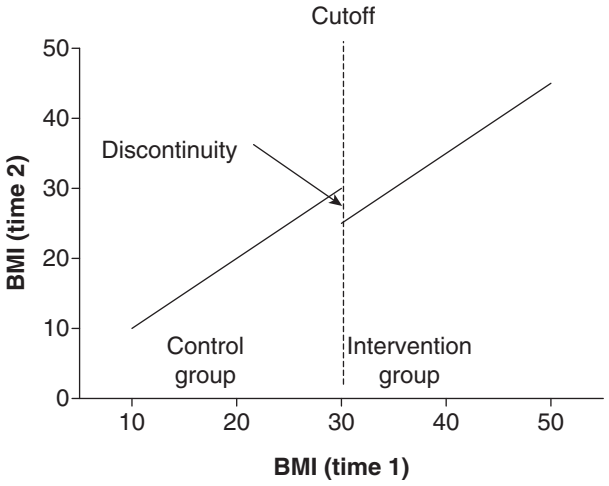



**Figure 1** Pre- Versus Postintervention Body Mass Indexes (BMIs) for Treated and Control Participants in a Weight-Loss Program

based on the pretest. For example, the pretreatment and cutoff scores could be the BMI, and the posttest score could be caloric intake. Similarly, caloric consumption could be measured before and after the intervention, but the cutoff could be based on the BMI; or caloric intake could be measured at Time 1, weight measured at Time 2, and BMI used to determine the cut point.

### Advantages

The gold-standard design for evaluating the effectiveness or efficacy of an intervention is the randomized controlled trial (RCT). However, there is an ethical problem with RCTs. If the intervention works, then half the participants will have been denied it (at least until the study ends). In contrast, all eligible people (or groups) in an RD study receive the treatment, to the degree that resources permit.

A second advantage is that if an RD design is correctly implemented and analyzed, the results are an unbiased estimate of the treatment effect. Moreover, the inclusion of a comparison group is protection against many of the threats to internal validity found in many other quasi-experimental techniques. For example, if there were external factors that could modify the relationship between the pre- and postmeasures, then these should act in a similar fashion in both groups (except for the highly unlikely circumstance that the factors operate only at the exact cutoff point). Indeed, the internal validity of an RD study is as strong as that of an RCT.

Third, because the groups are compared only at the cut point, the RD design is not susceptible to the biases affecting other pre–post designs, such as maturational effects or regression to the mean; and it is not necessary to assume that the data are interval, which is the case when difference scores are used.

### Disadvantages

There are a number of disadvantages with the RD design. First, it tends to be less powerful than an RCT. This means that more participants must be enrolled in order to demonstrate a statistically significant effect for a given effect size. Some estimates are that 2.75 times as many people are needed in an RD design than in an equivalent RCT. If the unit of allocation is the group rather than the individual (e.g., a classroom or hospital), then the sample size determination must be adjusted to account for the correlation of scores within the group (i.e., clustering effects).

A second problem is a direct consequence of the lack of random assignment of people to groups. In an RCT, randomization tends to minimize group differences, so that any result can be attributed to the intervention. Because this is not the case in an RD design, there is an assumption that there is no discontinuity between the groups in terms of the relationship between the two measures that, just by accident, happens to coincide with the cutoff.

A third potential problem is that the validity of the results is dependent on the form of the relationship between the pre- and postmeasures. For example, if there is a quadratic (i.e., curvilinear) relationship, then fitting a straight line may lead to the erroneous conclusion that there is a discontinuity at the cutoff point, when in fact there is none.

*David L. Streiner*

***See also*** Applied Research; Experimental Design; Quasi-Experimental Design

#### Further Readings

Braden, J. P., & Bryant, T. J. (1990). Regression discontinuity designs: Applications for school psychology. *School Psychology Review, 19,* 232–239.

Shadish, W. R., Cook, T. D., & Campbell, D. E. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin.

Thistlewaite, D., & Campbell, D. E. (1960). Regression discontinuity analysis: An alternative to the ex-post-facto experiment. *Journal of Educational Psychology, 51,* 309–317.

Zuckerman, I. H., Lee, E., Wutoh, A. K., Xue, Z., & Stuart, B. (2006). Application of regression-discontinuity analysis in pharmaceutical health services research. *Health Services Research, 41,* 550–563.

## Regression to the Mean

Regression to the mean (RTM) is a widespread statistical phenomenon that occurs when

a nonrandom sample is selected from a population and the two variables of interest measured are imperfectly correlated. The smaller the correlation between these two variables, the more extreme the obtained value is from the population mean, and the larger the effect of RTM (that is, there is more opportunity or room for RTM).

If variables $X$ and $Y$ have standard deviations $SDx$ and $SDy$, and correlation $= r$, the slope of the familiar least squares regression line can be written $rSDy/SDx$. Thus a change of one standard deviation in $X$ is associated with a change of $r$ standard deviations in $Y$. Unless $X$ and $Y$ are perfectly linearly related, so that all the points lie along a straight line, $r$ is less than 1. For a given value of $X$, the predicted value of $Y$ is always fewer standard deviations from its mean than is $X$ from its mean. Because RTM will be in effect to some extent unless $r = 1$, it almost always occurs in practice.

As discussed by Donald Campbell and David Kenny, RTM does not depend on the assumption of linearity, the level of measurement of the variable (for example, the variable can be dichotomous), or measurement error. Given a less than perfect correlation between $X$ and $Y$, RTM is a mathematical necessity. Although it is not inherent in either biological data or psychological data, RTM has important predictive implications for both. In situations in which one has little information to make a judgment, often the best advice is to use the mean value as the prediction.

## History

An early example of RTM may be found in the work of Sir Francis Galton on heritability of height. He observed that tall parents tended to have somewhat shorter children than would be expected given their parents' extreme height. Seeking an empirical answer, Galton measured the height of 930 adult children and their parents and calculated the average height of the parents. He noted that when the average height of the parents was greater than the mean of the population, the children were shorter than their parents. Likewise, when the average height of the parents was shorter than the population mean, the children were taller than their parents. Galton called this phenomenon regression toward mediocrity; we now call it RTM. This is a statistical, not a genetic, phenomenon.

## Examples

### *Treatment Versus Nontreatment*

In general, among ill individuals, certain characteristics, whether physical or mental, such as high blood pressure or depressed mood, have been observed to deviate from the population mean. Thus, a treatment would be deemed effective when those treated show improvement on such measured indicators of illness at posttreatment (e.g., a lowering of high blood pressure or remission of or reduced severity of depressed mood). However, given that such characteristics deviate more from the population mean in ill individuals than in well individuals, this could be attributable in part to RTM. Moreover, it is likely that on a second observation, untreated individuals with high blood pressure or depressed mood also will show some improvement owing to RTM. It also is probable that individuals designated as within the normal range of blood pressure or mood at first observation will be somewhat less normal at a second observation, also due in part to RTM. In order to identify true treatment effects, it is important to assess an untreated group of similar individuals or a group of similar individuals in an alternative treatment in order to adjust for the effect of RTM.

### *Variations Within Single Groups*

Within groups of individuals with a specific illness or disorder, symptom levels may range from mild to severe. Clinicians sometimes yield to the temptation of treating or trying out new treatments on patients who are the most ill. Such patients, whose symptoms are indicative of characteristics farthest from the population mean or normality, often respond more strongly to treatment than do patients with milder or moderate levels of the disorder. Caution should be exercised before interpreting the degree of treatment effectiveness for severely ill patients (who are, in effect, a nonrandom group from the population of ill individuals) because of the probability of RTM. It is important to separate genuine treatment effects from RTM effects; this is best done by employing

randomized control groups that include individuals with varying levels of illness severity and normality.

### How to Deal With Regression to the Mean

If subjects are randomly allocated to comparison groups, the responses from all groups should be equally affected by RTM. With placebo and treatment groups, the mean change in the placebo group provides an estimate of the change caused by RTM (plus any other placebo effect). The difference between the mean change in the treatment group and the mean change in the placebo group is then the estimate of the treatment effect after adjusting for RTM. C. E. Davis and others (e.g., P. L. Yudkin & I. M. Stratton) showed that RTM can be reduced by basing the selection of individuals on the average of several measurements instead of a single measurement. They also suggested selecting patients on the basis of one measurement but to use a second pretreatment measurement as the baseline from which to compute the change. It was noted that if the correlation coefficient between the posttreatment and the first pretreatment measurement is the same as that between the first and the second pretreatment measurement, then there will be no expected mean change due to RTM. Of course, understanding the phenomenon of RTM is the first step to overcoming the problems caused by RTM.

*Sophie Chen and Henian Chen*

***See also*** Distribution; Mean; Variable

#### Further Readings

Bland, J. M., & Altman, D. G. (1994). Some examples of regression towards the mean. *British Medical Journal, 309,* 780.

Campbell, D. T., & Kenny, D. A. (1999). *A primer of regression artifacts.* New York: Guilford.

Davis, C. E. (1976). The effect of regression to the mean in epidemiologic and clinical studies. *American Journal of Epidemiology, 104,* 493–498.

Galton, F. (1886). Regression toward mediocrity in hereditary stature. *Journal of the Anthropological Institute of Great Britain and Ireland, 15,* 246–263.

Yudkin, P. L., & Stratton, I. M. (1996). How to deal with regression to the mean in intervention studies. *Lancet, 347,* 241–243.

## RELIABILITY

This entry focuses on reliability, which is a desired property of the scores obtained from measurement instruments. Therefore, reliability is relevant in a variety of contexts, such as scores on multiple-choice questions on an achievement test and Likert-type scale responses on a survey questionnaire. The concept of reliability is premised on the idea that observed responses are imperfect representations of an unobserved, hypothesized latent variable. In the social sciences, it is typically this unobserved characteristic, rather than the observed responses themselves, that is of interest to researchers. Specifically, reliability serves to quantify the precision of measurement instruments over numerous consistent administration conditions or replications and, thus, the trustworthiness of the scores produced with the instrument. This entry discusses several frameworks for estimating reliability.

### Reliability as Replication

While it is assumed that items (also called questions, tasks, prompts, or stimuli) on an instrument measure the same theoretical construct, this assumption always needs to be tested with empirical data. Reliability does not provide direct information about the meaningfulness of score interpretations, which is the concern of validation studies, but serves as an empirical prerequisite to validity.

Reliability provides information on the replicability of observed scores from instruments. Consequently, at least two sets of scores are required to obtain information about reliability. These can be multiple administrations of an instrument to the same group of respondents (*test–retest reliability* or *score stability*), the administration of two comparable versions of the instrument to the same group of respondents (*parallel-forms reliability*), or the comparison of scores from (at least) two random halves of an instrument (*split-half reliability* and *internal consistency*).

To estimate test–retest reliability, the original instrument has to be administered twice. The length of the time interval is not prespecified but should be chosen to allow variance in performance without the undue influence of developmental

change; ideally, the time interval should be reported along with the reliability estimate. To estimate parallel-forms reliability, two forms of the instrument have to be developed according to the same set of test specifications and have to be administered to the same group of respondents. As with test–retest reliability, the time interval between administrations is important and should be reported.

To estimate the split-half reliability or internal consistency of an instrument, only one instrument needs to be administered to one group of respondents at one point in time. Technically, the instrument needs to be split into at least two randomly parallel halves. Of course, there are many ways to divide the items on the instrument into two halves. To overcome the limitation imposed by the arbitrariness of the split, coefficient alpha has been developed. It is theoretically equivalent to the average of all potential split-half reliability estimates and is easily computed. Estimating the internal consistency of an instrument effectively measures the homogeneity of the items on the instrument.

## Reliability as a Theoretical Quantity

To understand the definitions of different reliability coefficients, it is necessary to understand the basic structure of the measurement framework of classical test theory (CTT). In CTT, observed scores ($X$) are decomposed into two unobserved components, true score ($T$) and error ($E$; i.e., $X = T + E$). While true score is assumed to be stable across replications and observations, error represents a random component that is uncorrelated with true score. Additional restrictions can be placed on the means and variances of the individual components as well as their correlations in order to make them identified and estimable in practice.

In particular, using a few basic assumptions, true score variance can be defined as the sum of observed score variance and error variance (i.e., $Var(X) = Var(T) + Var(E)$). Reliability is simply the amount of variation in the observed scores that is due to true interindividual differences. Alternatively, the reliability coefficient can be viewed as the ratio of signal to signal-plus-noise. More technically, it is the ratio of true score variance to observed score variance. Using $\rho_{XX'}$ to denote the reliability coefficient, $\sigma_X^2$ to denote observed score variance, $\sigma_T^2$ to denote true score variance, and $\sigma_E^2$ to denote error variance, the reliability of scores for item $X$ is defined as

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2}.$$

The value of this reliability coefficient tends to 0 when the variance of the true scores in the population goes to 0 (i.e., when there is little interindividual variation in the population). The reliability coefficient will approach 1 if the variance of the true scores in the population goes to infinity (i.e., if there is very little measurement error relative to the amount of interindividual differences in the population).

## Estimates of Reliability

It is important to note that the reliability coefficient defined above is a theoretical quantity that is always positive because variance components are always positive. Moreover, it cannot be estimated directly, because true score variation is not directly observable from data. Instead, the true reliability of scores has to be estimated from data using suitable sample coefficients. Essentially, most of the sample coefficients that are used to estimate the true reliability of scores are *correlation coefficients*.

The most common coefficient in practice is the Pearson product-moment correlation coefficient (PPMCC). To estimate test–retest reliability, the scores on the same instrument from the two measurement occasions are correlated by means of the PPMCC; to estimate parallel-forms reliability, the scores from the two forms of the instrument from the two measurement occasions are correlated by means of the PPMCC; and to estimate split-half reliability, the scores from two randomly selected halves are correlated by means of the PPMCC.

### *The Spearman–Brown Formula*

Generally speaking, reliability increases when homogeneous items are added to an existing instrument. Homogeneous items are items whose correlations with other items on the augmented form equal the average interitem correlation of the existing form. This principle can be used in estimating the reliability of shortened or lengthened

instruments by using what is known as the Spearman–Brown Prophecy formula. This formula is defined as

$$_{SB}\rho_{XX'} = \frac{n\rho_{XX'}}{[1 + (n-1)\rho_{XX'}]},$$

where $n$ is an integer value representing the ratio of projected form length to original form length. To estimate the reliability of a full form via a split-half reliability coefficient, this simplifies to

$$_{SB}\rho_{XX'} = \frac{2\rho_{X_1X_2}}{1 + \rho_{X_1X_2}},$$

where $\rho_{X_1X_2}$ is the correlation between the observed scores on the two halves.

### *The Kuder–Richardson Formulas*

Alternatively, for instruments comprising dichotomously scored items solely, two often cited reliability coefficients are the Kuder–Richardson (KR) coefficients KR-20 and KR-21. The KR-20 reliability coefficient is estimated as

$$_{20}\rho_{XX'} = \left(\frac{n}{n-1}\right)\left(1 - \frac{\sum_{all\ items} P(1-P)}{\sigma_X^2}\right),$$

where $n$ is the number of items on the form, $P$ is the proportion of respondents getting a dichotomously scored item correct, and $P(1 - P)$ is simply the resulting variance of that item. KR-21 makes the additional restrictive assumption that all items have the same proportion correct, allowing the reliability coefficient to be calculated when only number of items, average score, and standard deviation of observed scores are known. It is computed as

$$_{21}\rho_{XX'} = \left(\frac{n}{n-1}\right)\left(1 - \frac{\mu_X(n-\mu_X)}{n\sigma_X^2}\right),$$

where $\mu_X$ is the mean observed score on the form.

### *Cronbach's α*

Perhaps the most well-known internal consistency reliability coefficient is Cronbach's alpha, which assumes, in its basic form, that the total score on an instrument is computed as the simple (i.e., unweighted) sum of the scores from all items. Cronbach's alpha is a generalization of the KR-20 coefficient and is applicable to dichotomously scored items as well as polytomously scored items. It is defined as

$$_{\alpha}\rho_{XX'} = \left(\frac{n}{n-1}\right)\left(\frac{\sigma_X^2 - \sum_{all\ items}\sigma_{X_I}^2}{\sigma_X^2}\right)$$

$$= \left(\frac{n}{n-1}\right)\left(1 - \frac{\sum_{all\ items}\sigma_{X_I}^2}{\sigma_X^2}\right),$$

where $\sigma_X^2$ is the observed score variance for the total scores and $\sigma_{X_I}^2$ is the observed score variance for the item scores.

## Reliability and Standard Error

High degrees of reliability for scores are, of course, desirable. The more reliable scores on an instrument are, the less random error there is in the observed responses, and thus, the more precise the information about respondents will be. While reliability coefficients are quantities that index in a single number the precision of an instrument, it is often equally important to quantify the resulting error that accompanies latent variable estimates for respondents as a result of the imperfection in measurement.

There are essentially two types of error that are relevant in a CTT model, (1) *unconditional standard error of measurement* (*SEM*) and (2) *conditional SEM*. The former is a single number that is assumed to be the same across all values of the latent variable (i.e., it is the same for all respondents), and the latter is different across different values of the latent variable (i.e., is larger for some respondents and smaller for others). Unconditional *SEM* is simply the standard deviation of the error scores and can be shown to be $\sigma_E = \sigma_X^2\sqrt{1 - \rho_{XX'}}$. As the formula shows, a perfect reliability implies no standard error (i.e., the observed scores are the true scores), while no reliability implies that all observed variation is simply error variation (i.e., the observed scores contain no information about interindividual differences). Note that an estimate of the true reliability has to be inserted into this formula, and different estimates will produce different values of *SEM*.

**Table 1** Relationship of the Population Values of Selected Internal Consistency Coefficients to the True Reliability

| | *Spearman–Brown* | *KR-20* | *KR-21* | *Rulon* | *Flanagan* | *Coefficient α* |
|---|---|---|---|---|---|---|
| SP | = | = | = | = | = | = |
| E-SP | = | = | = | = | = | = |
| TE | ≤ | = | = | = | = | = |
| E-TE | ≤ | = | ≤ | = | = | = |
| C | ≤ | ≤ | ≤ | ≤ | ≤ | ≤ |
| E-C | ≤ | ≤ | ≤ | ≤ | ≤ | ≤ |

*Note:* SP = strictly parallel; E-SP = essentially strictly parallel; TE = tau equivalent; E-TE = essentially tau equivalent; C = congeneric; E-C = essentially congeneric; KR = Kuder–Richardson.

Conditional *SEM* is, technically, the more appropriate measurement of error to use in practice. Essentially, error always varies across the range of latent variable values, and using the *SEM* is an unnecessary oversimplification. Different approaches can be taken to estimate conditional *SEM* under a CTT framework, including a binomial model for error, a normal model for error, and a combined model. Once a suitable model has been chosen, the conditional *SEM* can be computed and can be used to compute confidence intervals for the true latent variable scores of respondents. These intervals will differ in width at different points on the score continuum.

## Properties of Reliability Coefficients

It is important to realize that the formulas for the coefficients presented so far include population values for variances. In practice, these coefficients are estimated on the basis of sample data. As a result, one needs to distinguish between (a) the sample estimate of the reliability coefficients, (b) the population value of the reliability coefficients, and (c) the theoretical value of the reliability coefficient that these are trying to estimate. This reliance on sample data has two key implications.

The first implication is that these coefficients estimate the true reliability differentially well depending on the relationship that exists between scores under a CTT framework. Simply put, there are three different types of score relationships that form a decreasingly restrictive hierarchy in terms of the assumptions that are made about terms in the CTT model. The three relationships are (1a) strictly parallel, (2a) tau equivalent, and (3a) congeneric test scores. If mean differences across forms are further taken into account, one can further distinguish between (1b) essentially strictly parallel, (2b) essentially tau equivalent, and (3b) essentially congeneric test scores. All these models for test score sets can be tested using methods from factor analysis with mean structures.

Scores from strictly parallel items or forms of an instrument, which have to be constructed according to identical sets of specifications or blueprints, possess identical statistical properties. From a factor-analytic perspective, this means that the loadings and error variances are identical across strictly parallel items or forms, which implies an equal level of reliability for the different items or forms. The requirements of strict parallelism are unlikely to hold in most practical situations; therefore, the following weaker forms have been defined. Scores from items or forms described as being tau equivalent are proportional to one another. From a factor-analytic perspective, this means that the loadings are still identical across tau-equivalent items or forms, but that the error variances are allowed to differ. Scores from congeneric items or forms assume only that the true scores are linearly related to one another. From a factor-analytic perspective, this means that the loadings and error variances are allowed to differ across congeneric items or forms. If mean differences across forms are further modeled, this leads to the essentially strictly parallel, essentially tau equivalent, and essentially congeneric items or forms, as introduced above. From a factor-analytic perspective, this implies that the items or forms can also differ in terms of their intercepts.

Table 1 lists selected properties of selected coefficients to illustrate how these different score relationships impact the relationship between the population values of the coefficients and the true reliability that they are trying to estimate. For

example, it shows that the population value of Cronbach's $\alpha$ equals the true reliability if the scores on the items on which it is computed are at least essentially tau equivalent. It is important to note, however, that the sample value of the coefficient can still under- or overestimate the population value and, thus, the true reliability due to random sampling error.

Thus, it is necessary to test which of the CTT models holds for a given instrument by using sample data to be able to gauge how well a reliability coefficient is likely to estimate the true reliability.

The second implication is that sample values of reliability coefficients should be used to construct confidence intervals and to conduct hypothesis tests to make inferences about the population values of the coefficients according to the appropriate statistical theory. For example, one can construct confidence intervals for the population value of Cronbach's $\alpha$ in a single population as well as across multiple populations.

## Reliability in Alternative Measurement Frameworks

The previous discussion has focused on reliability estimation in a CTT framework, which is probably the most common measurement framework that is used in practice. However, alternative frameworks that provide definitions of reliability coefficients are available for modeling variation in latent variables. This section briefly discusses generalizability theory (g theory), factor analysis (FA), and item response theory (IRT).

### *G Theory*

G theory breaks up the error component in the basic CTT model to differentiate multiple sources of error. To estimate the relative contributions of these different error sources, which are viewed as *factors* or *facets* from an experimental design perspective, random-effects analysis of variance procedures are used. Depending on the number of factors that are used in the design and depending on whether they are experimentally crossed with one another or nested within one another, different types of reliability coefficients can be defined and estimated.

There are two primary types of reliability coefficients in the g-theory framework, namely the generalizability coefficient and the dependability coefficient. The former is analogous to a traditional CTT reliability coefficient and is calculated as the ratio of true or universe score variance to the expected observed score variance. As in CTT, this coefficient is designed to quantify the uncertainty associated with norm-referenced or relative comparisons of respondents based on scores. The latter estimates the precision of an observed score in estimating a person's true score. Thus, unlike in traditional CTT models, this coefficient is designed to quantify the uncertainty associated with criterion-referenced or absolute decisions.

Mathematically, both coefficients can be defined as

$$\Phi = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_E^2},$$

which resembles the formula for the reliability coefficient in the CTT framework. However, for the generalizability coefficient, $\sigma_E^2$ is composed of the variance of items ($\sigma_I^2$) only, whereas for the dependability coefficient, $\sigma_E^2$ is composed of the variance of the items and the person-by-item interaction ($\sigma_I^2 + \sigma_{pI}^2$). As a result, the error variance for the dependability coefficient is always greater than the error variance of the generalizability coefficient.

### *Factor Analysis (FA)*

Factor-analytic techniques can be employed to investigate the score relationships for items on a single instrument or for items on multiple instruments. Specifically, FA can be used to empirically test which of the six basic CTT models is likely to hold in the population. This technique involves the use of sample data with appropriate restrictions on factor loadings, error variances, and latent means. The reliability for an instrument consisting of multiple items can be estimated as

$${}_{FA}\rho_{XX'} = \frac{\sum\limits_{all\ items} \lambda_I^2}{\sum\limits_{all\ items} \lambda_I^2 + \sum\limits_{all\ items} \sigma_E^2},$$

where $\lambda_I^2$ is the factor loading associated with each item and $\sigma_E^2$ is the error variance associated with each item. Depending on whether factor loadings,

error variances, or both are constrained to equality across items, reliability estimates for different CTT models can be computed.

Mean-structure models can also be fit, and the resulting models can be compared to models without mean structures, but the mean structures do not have an effect on the reliability estimate itself. However, testing for mean structure is important in order to find which CTT model is likely to hold in the population, which helps to explain the statistical properties of different reliability estimators. In general, FA is a powerful framework for estimating the reliability coefficient that is most suitable for the data structure at hand if the sample size is sufficiently large to provide stable estimates of loadings, error variances, and latent means.

### *IRT*

IRT uses a nonlinear function to model the relationship of observed scores to true scores at the item level. Put simply, IRT is conceptually similar to a simple logistic regression model with a latent predictor variable, which gets estimated simultaneously across all items for an instrument. There are close relationships between some basic IRT and CTT models and, thus, between the concept of reliability in IRT and CTT.

More specifically, reliability in IRT is marginally defined as the average reliability across all potential values of the latent variable. Of greater interest than a single reliability coefficient in IRT are the information function and the standard error of estimation (*SEE*) function, however. The former shows how much statistical information each item provides for each value of the latent variable (i.e., for different respondents) while the latter shows how much uncertainty about the latent variable value there is for different respondents.

Just as with the relationship of reliability with unconditional and conditional *SEM*, there is a close relationship between reliability and the statistical information provided via *SEE* in IRT. Essentially, the statistical information at different latent variable values is inversely proportional to the *SEE* such that items that provide a lot of information about certain respondents have a small *SEE* for them (i.e., provide very precise estimates for them). Moreover, just as conditional *SEM* should always be used in a CTT framework to quantify the precision of latent variable estimates for respondents, *SEE* should always be used in an IRT framework in addition to simply computing reliability estimates.

*Matthew M. Gushta and André A. Rupp*

***See also*** Classical Test Theory; Coefficient Alpha; Generalizability Theory; Item Response Theory; Replication; Spearman–Brown Prophecy Formula; Standard Error of Measurement; Structural Equation Modeling; Validity of Measurement

### Further Readings

Brennan, R. L. (2001). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement, 38*(4), 295–317.

Brennan, R. L. (2001). *Generalizability theory.* New York: Springer.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78,* 98–104.

Doran, H. C. (2005). The information function for the one-parameter logistic model: Is it reliability? *Educational & Psychological Measurement, 65*(5), 665–675.

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability. *Educational & Psychological Measurement, 66,* 930–944.

Haertel, E. (2006). Reliability. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65–110). Westport, CT: Praeger.

Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. *Psychological Methods, 1,* 293–299.

Miller, M. B. (1995). Coefficient alpha: A basic introduction from the perspectives of classical test theory and structural equation modeling. *Structural Equation Modeling, 2*(3), 255–273.

Millsap, R. E., & Everson, H. E. (1991). Confirmatory measurement model comparisons using latent means. *Multivariate Behavioral Research, 26,* 479–497.

Raykov, T. (1997). Scale reliability, Cronbach's coefficient alpha, and violations of essential tau-equivalence with fixed congeneric components. *Multivariate Behavioral Research, 32,* 329–353.

Reuterberg, S., & Gustafsson, J. (1992). Confirmatory factor analysis and reliability: Testing measurement model assumptions. *Educational and Psychological Measurement, 52,* 795–811.

Traub, R. E., & Rowley, G. L. (1991). Understanding reliability. *Educational Measurement: Issues & Practice, 10,* 37–45.

# Repeated Measures Design

Repeated measures experiments are appropriate when multiple measures of a dependent variable are taken on the same subjects or objects or matched subjects or objects under different conditions or over two or more time periods. Repeated measures cover a broad range of research models, from comparison of two treatments on the same subjects to comparisons of multiple treatments on multiple levels of two or more factors on the same subjects and to assessing the differences in means among several related scales that share the same system of measurement.

The terminology used for repeated measures varies among researchers and statisticians. For example, repeated measures designs are also known as repeated measures analysis of variance (ANOVA), repeated measures factors, or within-subjects designs. However, repeated measures may be performed within a multivariate ANOVA design as well. The logic of repeated measures is based on the calculation of difference scores (contrasts) for comparisons of treatment levels for each subject. For example, if there were no treatment effect, the difference between scores before and after the treatment would be close to zero. The number of contrasts is usually 1 less than the number of levels in a factor. A factor with two levels would produce one difference score whereas a factor with three levels would produce two difference scores: one score for the difference between Level 1 and Level 2, and one score for the difference between Level 2 and Level 3, unless other specific comparisons are requested. Contrasts are pooled together to provide the significance test. The mean of contrast scores and variation around the sample mean difference based on the individual difference scores are used to determine whether there is a significant difference in the population.

Repeated measures designs include a broad class of statistical models with fixed effects and/or random effects. Fixed effects are changes in the means of the dependent variable that can be attributed to the independent variables. For example, differences in mean attitude scores about the environment between individuals who live in a rural location and those in an urban location may be attributed to the fixed effect of location. Location is a fixed factor because the researcher made a conscious decision to include only rural and urban locations, although other locations such as suburban could have been included. Random effects are changes in the variance of the dependent variable that can be attributed to the sampling units. For example, participants in a study are usually considered to be a random factor because the researcher selects a sample from a larger population. If another sample were selected, results would likely be different.

Repeated measures designs differ from the traditional between-groups designs that require a separate sample of subjects for each measurement or treatment condition. Observations in a repeated measures design may be from the same sample or experimental unit from one time to the next or from one condition to the next. In addition, repeated measures may be taken on matched pairs of individuals from different groups. Each individual in one of the groups is matched, on the specific variable being investigated, with a similar individual in another group. Measures may be taken over specified or unspecified time periods. Measures taken in a repeated measures study cannot be considered independent of one another because the measures are taken on the same or matched subjects, and one would expect that such measures would be correlated.

## Applications

Repeated measures analyses are used for studies in many different disciplines, ranging from behavioral and social sciences to medicine, engineering, and agriculture. Following are examples of research problems that are appropriate for repeated measures procedures: effects of three different treatments (chemicals or fertilizers) on two different kinds of crops; changes in behavior before and after an intervention; difference among scores on several related scales; reaction time under four different conditions; student performance before and after an intervention; performance of different machines at specified time periods; reactions to drugs or treatments over time; and ratings on specific products or services.

Repeated measures may be taken at two or more different times. For example, energy level of a sample may be taken at 7:00 a.m., 1:00 p.m.,

and again at 6:00 p.m. Measures taken at different times can vary based on amount and quality of sleep, type of breakfast that the individual ate, type of activity performed before each measure was taken, and so on. Measures taken on one group and compared with measures from a different group may be invalid because individual activities, habits, and natural body clocks vary. Repeated measures on the same subjects allow the subjects to act as their own control.

Repeated measures may be taken under two or more different conditions. When only two measures are taken, the design may be referred to as a paired-test design, as in a study comparing anxiety when measures are taken on the same subject tested in two different situations: high-stakes testing and testing with little or no consequences. The sample means for each testing condition are examined to assess statistically significant differences between the two testing conditions in the population.

## Types

### *Paired Samples* t *Test*

The paired samples *t* test is a basic research design in terms of convenience and analysis. This test is also known as a correlated samples design, and it is the simplest of the within-subjects designs. The purpose is to compare mean scores within subjects by testing the null hypothesis of no difference between the two measures. Paired samples *t* tests are performed on one sample in which each participant in the sample is measured two times on the dependent variable. This is the case in the pretest–posttest situation in which each participant is measured before a treatment and again after the treatment. For example, measuring college students' knowledge about the environment before a special unit of instruction is taught and again after the unit is taught is a paired samples *t* test design. Participants serve as their own control because both measures are taken on the same participants. Paired comparisons are also appropriate in cases in which repeated measures are not collected from the same participants. Participants may be matched through a natural pairing process, such as sisters and brothers, or through a testing process in which participants are matched on the basis of test scores.

The paired samples *t* test should meet certain statistical assumptions; for instance, the data for each variable should be from a randomly selected and normally distributed population, individual scores within a group should be independent of one another, and population variances should be equal. The assumption of equal variances is known as homogeneity of variance. The null hypothesis of equal means is rejected or retained based on the *t* value at the associated degrees of freedom and *p* value. If there is a low probability that the results occurred by chance at the preset alpha level (generally .05), the null hypothesis is rejected; otherwise it is retained—never accepted.

### *One-Way Repeated Measures ANOVA*

The one-way repeated measures ANOVA is an extension of the paired samples *t* test to more than two levels of a factor. The purpose of a one-way repeated measures ANOVA is to determine whether there are statistically significant differences between or among the means associated with different levels of a factor. The null hypothesis states that there are no differences among the levels of a factor ($\mu_1 = \mu_2 = \cdots = \mu_n$ ). The design involves a treatment factor and a time or condition factor. Subjects are matched on one or more variables, with random assignment of subjects to levels of the independent variable. The number of treatments is set by the researcher. The order in which treatments are processed should be random if order could affect the results. The treatment on the same subject represents the between-subjects factor, and the time or condition factor for observations at different times or under different conditions on the same subject represents the within-subjects factor.

Statistical assumptions for the one-way repeated measures ANOVA are similar to those of the paired samples *t* test with two additions: treatment levels are independent of one another, and the covariance matrix for the orthonormalized contrasts have equal variances and covariances for all levels of a factor (compound symmetry). Equal covariances among levels of a factor are known as homogeneity of covariance or sphericity. The assumptions of compound symmetry and sphericity must be met in order for results of the statistical tests to be interpreted accurately. Violations of the

homogeneity-of-variance assumption are more serious when group sizes are unequal.

One-way repeated measures ANOVA is also known as a *completely randomized design* if all subjects are assigned randomly to the treatments; therefore, there are no blocking effects. The absence of blocking poses problems in interpreting the results if the sampling units are not homogeneous. Repeated measures designs are usually balanced designs in that an equal number of subjects is included at each level of the independent variable. Null hypotheses are rejected or retained on the basis of multivariate test results, which are usually reported to avoid a possible violation of the sphericity assumption. Along with the $F$ value, associated degrees of freedom, and $p$ value, the Wilks's lambda is the most commonly reported multivariate test result for the one-way repeated measures ANOVA. For factors with more than two levels and significant multivariate test results, paired samples $t$ tests or polynomial contrasts should be conducted to identify which levels are different from one another.

### *Two-Way Repeated Measures ANOVA*

The two-way repeated measures ANOVA design involves two independent variables (factors) in the analysis, with two or more levels each. The two-way repeated measures ANOVA is sometimes referred to as a *factorial* ANOVA or *two-factor* ANOVA. The purpose of the two-way repeated measures ANOVA is to determine whether there is a significant main effect of each of the factors on the dependent variable and an interaction effect between the two factors. Three null hypotheses are tested: one for each main effect and one for the interaction effect. Each subject is tested at each level of both factors. Adding a second factor, such as time, with three levels, to an experiment with a condition factor with two levels would be considered a two-way design, as long as repeated measures are taken on the same or matched subjects. The addition of a control group with repeated measures on the experimental group and the control group would be a two-way repeated measures ANOVA design. If the repeated measures are taken on only one factor, the design is called a *mixed-model* ANOVA.

Assumptions for the two-way repeated measures take into account the two factors in the analysis. Like the one-way repeated measures ANOVA, the assumptions of normality, sphericity, and random selection of subjects should be met. In addition, the difference scores for the multivariate tests should be multivariately normally distributed, and the difference scores from one subject to the next should be independent.

Like the one-way repeated measures procedure, the results of the two-way repeated measures ANOVA are usually interpreted on the basis of the multivariate tests to avoid violations of the sphericity assumption. Each factor and the interaction of the factors are listed in the multivariate table with their corresponding test results. Null hypotheses are rejected or retained based on Wilks's lambda, the $F$ value, associated degrees of freedom, and the $p$ value. Wilks's lambda is the most commonly reported multivariate test for main effects and interaction effects. Post hoc follow-up tests should be conducted for significant interaction effects. If there is no significant interaction, but a significant main effect, follow-up tests should be conducted on factors with more than two levels to identify which levels are different from one another.

### *Two-Way Mixed Model ANOVA*

Two-way mixed model ANOVA designs test differences within groups with repeated measures, as well as differences between groups. Mixed models are also known as *two-factor between-within* ANOVAs or *split-plot* ANOVAs.

The major difference between the two-way repeated measures ANOVA and the mixed model ANOVA is that in a two-way repeated measures design, each subject is tested at each level of two different factors whereas in the mixed model ANOVA, repeated measures are made on only one of the two factors. For example, a study of environmental concerns between rural and urban citizens before and after a community awareness program would be important to sociologists. In this study, geographic location (rural or urban) is a between-groups factor, and the within-subjects (repeated measures) factor is pre–post measures of community awareness. The main effects of geographic location and the interactions of location

with the pre–post measures can be examined. Another example is a single treatment applied to a large experimental unit, with multiple treatments administered randomly to subunits of the larger experimental unit. The between-subjects factor is derived from the larger unit, and the within-subjects factor is based on the treatments to the subunits. A more complex example of the between–within design is the random assignment of four subjects to three different conditions under which four different measures are taken on a variable of interest. The three conditions represent the between-factor variable and the four measures are the repeated factor.

Statistical assumptions for mixed models are more complex than for the one-way and two-way situations because assumptions for both between- and within-subjects analyses need to be considered. In addition, the variance and covariance structures are more complex. The assumptions of normality and independence should be satisfied for both between- and within- groups. The assumption of homogeneity of covariance should be met for between-groups analysis, and the assumption of sphericity should be met for within-groups analysis. Null hypotheses should be rejected or retained based on results of the statistical tests. Wilks's lambda is the most commonly reported multivariate test for main effects and the interaction effect, along with the $F$ value, associated degrees of freedom, and the $p$ value. Post hoc follow-up tests should be conducted for significant interaction effects. If there is no significant interaction, but a significant main effect, follow-up tests should be conducted on factors with more than two levels to identify which levels are different from one another.

### *Latin Square Designs*

The Latin square design counterbalances the order of treatments to ensure that each treatment condition occurs equally often in the experiment. The purpose is to compare the effects of repeated observations on multiple levels of a factor for multiple sampling units. Latin square designs include three factors, which may be visualized as levels of one factor (sampling units) represented on the rows of a square matrix, levels of another factor (conditions or time periods) on the columns of the matrix, and a third factor represented by the intersections of the rows and columns. Each factor must have the same number of levels, resulting in a square matrix, thus a Latin square design. Measures for the combinations of factors on the rows and columns are the treatment effects. All combinations of the factor levels may or may not be included in the design. Latin square designs are similar to nested designs, which offer control over the variation introduced by the sampling units. For example a study of consumer ratings of the usability of a specific product under four different conditions throughout the year could be structured as a Latin square design with the required three factors: four consumers listed on the rows, four conditions or time periods listed on the columns, and consumer ratings (treatment effects) listed in the row and column intersections. Assumptions for the Latin square design are the same as for other repeated measures designs. Trend analyses are recommended in cases in which order of treatments is important or practice or carry-over effects may be present. Results of a Latin square analysis are displayed in an ANOVA table.

## Advantages and Disadvantages

### *Advantages*

A major advantage of a repeated measures design is that subjects are used as their own control because each subject is a member of the control group and the experimental group. Using the same participants in both groups can help cut the cost of a study, especially in situations in which participants are difficult to find. A measure taken on one individual and compared with the same measure on another individual does not give valid information on variability due to within-subject difference.

### *Disadvantages*

The greatest disadvantage with repeated measures designs is possible carryover effects. Problems can arise with subject fatigue, boredom, or practice effects that carry over from one measure to the next. These problems can be minimized by increasing the time between measures and randomizing the order in which treatments are

administered. There may also be unobserved or hidden effects that are not discovered unless a second treatment is introduced, which may not be feasible.

*Marie Kraska*

***See also*** Analysis of Variance (ANOVA); Block Design; Latin Square Design; Pairwise Comparisons; *p* Value; Significance Level, Concept of; Significance Level, Interpretation and Construction; Split-Plot Factorial Design; Within-Subjects Design

**Further Readings**

Cardinal, R. N., & Aitken, M. R. F. (2006). *ANOVA for the behavioral sciences researcher.* Mahwah, NJ: Lawrence Erlbaum.

Everitt, B. S. (1995). The analysis of repeated measures: A practical review with examples. *The Statistician, 44*(1), 113–135.

Green, S. B., & Salkind, N. J. (2008). *Using SPSS for Windows and Macintosh: Analyzing and understanding data* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Hirotsu, C. (1991). An approach to comparing treatments based on repeated measures. *Biometrika, 78*(3), 583–594.

Looney, S. W., & Stanley, W. B. (1989). Repeated measures analysis for two or more groups: Review and update. *American Statistician, 43*(4), 220–225.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear statistical models* (4th ed.). Boston: McGraw-Hill.

Ramsey, F. L., & Schafer, D. W. (2002). *The statistical sleuth* (2nd ed.). Pacific Grove, CA: Duxbury.

Rao, R. V. (1998). *Statistical research methods in the life sciences.* Boston: Duxbury.

Wallace, D., & Green, S. B. (2002). Analysis of repeated measures designs with linear mixed models. In D. S. Moskowitz & S. L. Hershberger (Eds.), *Modeling intraindividual variability with repeated measures data: Methods and applications* (pp. 103–134). Mahwah, NJ: Lawrence Erlbaum.

# REPLICATION

Any research conclusion is suspect if it is based on results that are obtained by chance. An indication of happenstance is a failure to obtain the same results when the original study is repeated. *Replication* refers to obtaining the same results when a previous study is conducted again in as similar a way as possible. The replication issue is discussed here in the context of choosing experimentally two theories.

The *decay theory* says that learning dissipates with time if one does not rehearse what has been learned. It implies that longer delays give rise to more forgetting. Experimenter *D* manipulates *delay of recall* to test it and obtains data consistent with the theory.

The *interference theory* says that forgetting occurs when subsequent learning interferes with earlier learning. It is expected that one forgets more if one has more additional learning. Experimenter *E* finds support for it by manipulating *amount of the additional learning*.

The onus is on both *D* and *E* to render it possible for themselves or any interested researcher to repeat their respective study by describing succinctly its experimental objective, specific to-be-tested hypothesis, method and materials used, and data collection, as well as analysis procedures.

A failure to replicate either study does not necessarily mean that there is no support for either theory. First, the description of the original study may not be as succinct as required. Second, there may be subtle variations in execution when the original procedure is repeated (e.g., the tone used in instructing the subjects). Third, the repeat study cannot (or may not) have all the features of the original study. For example, it might be impossible (or inadvisable) to employ the same group of participants in the repeat study. When another random sample is selected from the original population for the repeat study, an atypical sample may be selected by chance. Even if it were possible to employ the original participants, the participants themselves might have changed in ways that are relevant to the study.

Be that as it may, replicating *D*s (or *E*s) results proves neither the decay nor the interference theory. The variables, *delay of recall* and *amount of additional learning*, are confounded with each other in the sense that longer delays necessarily mean more additional learning; being given more additional learning may mean longer delays in testing the original learning. Hence, every successful replication of findings consistent with the decay theory is also consistent with the interference theory, and perhaps vice versa.

**Table 1** The Logic of Converging Operations in Theory Corroboration

| *The Theory of Interest* ($T_1$) *and Its Contenders* ($T_2$) | |
|---|---|
| **Theory $T_1$ and its implications ($I_{1j}$) in Situation $ABC_k$** | **A Contending theory ($T_2$) and its implication ($I_{2j}$) in Situation $ABC_k$** |
| If Theory $T_1$ is true, then $I_{11}$ in $ABC_1$. | If Theory $T_2$ is true, then $I_{21}$ in $ABC_1$. |
| If Theory $T_1$ is true, then $I_{12}$ in $ABC_2$. | If Theory $T_3$ is true, then $I_{32}$ in $ABC_2$. |
| If Theory $T_1$ is true, then $I_{13}$ in $ABC_3$. | If Theory $T_4$ is true, then $I_{43}$ in $ABC_3$. |
| If Theory $T_1$ is true, then $I_{14}$ in $ABC_4$. | If Theory $T_5$ is true, then $I_{54}$ in $ABC_4$. |
| If Theory $T_1$ is true, then $I_{15}$ in $ABC_5$. | If Theory $T_6$ is true, then $I_{65}$ in $ABC_5$. |

*Note:* $T_i$ = the $i$th theory; $I_{ij}$ = the $j$th implication of Theory $I_i$; $ABC_k$ = the $k$th situation.

In short, repeating an original study may duplicate a *confounding variable*. That is, replication does not disambiguate the confounding between two variables. This difficulty necessitates the adoption of *conceptual replication,* which is better categorized as *converging operations,* as a means of substantiating theories.

To avoid duplicating confounds (real or potential) of the original study, experimenters find it more instructive to use a series of theoretically informed and related experiments in which different independent variables or different experimental tasks may be used.

Any testable theory ($T_1$) should have well-defined implications (e.g., $I_{11}$ through $I_{15}$, as depicted in the left-hand panel of Table 1). The implicative relation is represented by the conditional proposition [CP], as shown in the left-hand panel of row 1 of Table 1.

> [CP]: If Theory $T_1$ is true, then Implication $I_{1j}$ should be observed in Condition $ABC_k$.

Relevant here are two rules of conditional syllogism. First, the *modus tollens* rule says that the antecedent of a conditional proposition (viz. $T_1$ in [CP]) is false if its consequent (i.e., $I_{1j}$) is not true. Hence, each of the implications of $T_1$ (viz. $I_{11}, \ldots, I_{15}$ in Table 1) is a criterion of rejection of $T_1$.

The second conditional syllogism rule is *affirming the consequent.* Knowing that a consequent (e.g., $I_{1j}$) is true does not guarantee that its antecedent ($T_1$) is true. It is for this reason that the tenability of $T_1$ is strengthened indirectly in the context of a contending theory (e.g., $T_2$ in the right-hand panel of row 1 in Table 1).

Not observing $I_{11}$ in Situation $ABC_1$ means that $T_1$ is false (by virtue of the modus tollens rule). Hence, setting up and collecting data in Situation $ABC_1$ is an explicit attempt to reject $T_1$. Observing $I_{11}$ (instead of $I_{21}$) in Situation $ABC_1$ leads to the rejection of $T_2$ but not a conclusive acceptance of $T_1$ in view of the affirming the consequent rule.

By the same token, observing $I_{12}$ (instead of $I_{32}$) in Situation $ABC_2$ leads to the rejection of $T_3$ but not a conclusive acceptance of $T_1$. Every failed attempt to reject $T_1$ (in Situations $ABC_3$, etc.) succeeds in rejecting a contending theory. At the same time, these attempts converge on the tenability of Theory $T_1$ from different angles. Hence, instead of appealing to replication, a series of theoretically informed converging operations should be used to substantiate a theory.

*Siu L. Chow*

***See also*** Experimental Design

**Further Readings**

Chow, S. L. (1992). *Research methods in psychology: A primer.* Calgary, Alberta, Canada: Detselig.

Garner, W. R., Hake, H. W., & Erikson, C. W. (1956). Operationalism and the concept of perception. *Psychological Review, 63,* 149–159.

# RESEARCH

Research is a form of organized intellectual work closely associated with the conduct of science. Like its cognates in French (*recherche*) and German (*Forschung*), the English word *research* carries connotations that distinguish it from such related terms as *discovery* and *scholarship*. This entry

explores the differences between research, discovery, and scholarship are explored, describes three crucial transition points from scholarship to research as a mode of inquiry, and offers some observations about research as a collective project of indefinite duration.

## Research Versus Discovery and Scholarship

The differences between the three terms can be grasped most easily through etymology.

*Discovery* suggests the revelation of something previously hidden, possibly forgotten, typically for a long time. This perhaps captures the oldest sense of knowledge in the Western intellectual tradition, traceable to the Greek *aletheia*. A discovery usually carries life-transforming significance because it purports to get at the original nature of things that are intimately connected to the inquirer's identity: We find out who we really are by learning where we belong in the great scheme of things. At the same time, though, in the words of Louis Pasteur, "Discovery favors the prepared mind," it is not clear that there is a method, let alone a logic, of discovery. Rather, discoveries are often portrayed as products of a receptive mind encountering a serendipitous event. The implication is that a discovery—as opposed to an invention—points to a reality that is never fully under the inquirer's control.

*Scholarship* suggests a familiarity with a wide range of established sources, the particular combination of which confers authority on the scholar. Scholarship thus tends to focus on the personality of the scholar, whose powers of discrimination are akin to those of a connoisseur who collects only the best works to perpetuate the values they represent. Unsurprisingly, the history of scholarship is intimately tied to that of editing and curation. Although some scholarship makes claims to originality, its value mainly rests on the perceived reliability of the scholar's judgment of sources. On this basis, others—especially students—may then study the scholar's words as accepted wisdom. Thus, the idea of scholarship already implies a public function, as if the scholar were conducting an inquiry on behalf of all humanity. This point is reflected in the traditional academic customs of annual public lectures and public defenses of doctoral dissertations.

*Research* suggests an exhaustive process of inquiry whereby a clearly defined field is made one's own. An apt metaphor here is the staking of a property claim bounded by the claims of other property holders. It is implied that a property holder is licensed to exploit the field for all the riches it contains. Indeed, research is strongly associated with the deployment of methods that economize on effort to allow for the greatest yield in knowledge. Unlike scholarship, research is by nature a private affair that is tied less to the intrinsic significance of what is investigated than to the effort invested in the activity. Consider it an application of the labor theory of value to the intellectual world. The transition between scholarship and research is most clearly marked in its public justification. The researcher typically needs to do something other than a version of his or her normal activities in order to demonstrate usefulness to others because of the inherently specialized nature of research.

Clearly, discovery, scholarship, and research are overlapping concepts for understanding the conduct of inquiry and the production of knowledge. For example, research can enhance the conditions of discovery by producing inquirers whose minds are sufficiently disciplined to recognize the true significance of an unanticipated phenomenon. For Thomas Kuhn, such recognition constitutes an "anomaly" for the researcher's "paradigm." Moreover, as the information scientist Don Swanson has observed, relatively little of published research is read, let alone cited. In effect, researchers unwittingly manufacture a vast textual archive of "undiscovered public knowledge" whose depths need to be plumbed by scholars who, equipped with smart search engines and other data-mining tools, may then discover solutions to problems that researchers in the relevant fields themselves believe still require further original work. Add to that Derek de Solla Price's observation that research activity has so rapidly intensified over the past half century that most of the researchers who have ever been alive are alive today, and one may conclude that, for better or worse, research is more in the business of manufacturing opportunities for knowledge than knowledge as such.

Implied in the above account is that discovery, scholarship, and research may correspond to

specific phases in the life cycle of scientific activity. In that case, *discovery* refers to the original prospective phase of epistemic insight, *research* to the ordinary work that Kuhn called normal science, and *scholarship* to the retrospective significance assigned to scientific work. The phenomenon of undiscovered public knowledge, which gives rise to the need for knowledge management, suggests that we may soon return to what had been the norm for doctoral-level research, even in the natural sciences, prior to the final third of the 19th century, namely, the testing of a theory against a body of texts. This reflects not only the surfeit of unread but potentially valuable written material but also the diminishing marginal rate of return on original research investment. Concerns about the unwittingly wasteful repetition of research effort had been already voiced in the 1950s, sparking the U.S. National Science Foundation to pilot what became the Science Citation Index, which researchers were meant to consult before requesting a grant to ensure that they were indeed heading in a new direction.

## Three Transition Points From Scholarship to Research

Three moments can be identified as marking the transition from scholarship to research as the leading form of organized inquiry in the modern era: the footnote, the doctoral dissertation, and the salary.

One significant trace is the *footnote*. Before the printing press enabled mass publishing, the growth of knowledge was slow, in large part because authors spent much effort literally reproducing the texts to which they then responded. As books became more easily available and journals circulated more widely, this literal reproduction was replaced by referencing conventions, starting with the footnote. Whereas the medieval scholastics had written their commentaries on the margins of the pages of books they copied, early modern scholars dedicated the main body of their texts to their own interpretations, while citing and contesting source works and those of other scholars in often lengthy footnotes. This practice may have left the impression that scholarship was always a messy and unresolved business, such that whatever was allowed to pass without criticism was presumed true, at least until further notice. The shift from a separate realm of footnotes to author-date citations internal to the main text corresponds to the emergence of a research mentality. In this context, other researchers and their sources are rarely contested. At most, the validity of their work is restricted in scope. But normally such work is cited in service of a positive case that the researcher wishes to make for a hypothesis or finding.

Another significant trace is the *doctoral dissertation*. In the older era of scholarship, a dissertation was seen primarily not as an original contribution to knowledge but as a glorified examination that provided candidates an opportunity to display their own brilliance in the mastery of sources, which included the dismissal of alternative interpretations. However, these dismissals were normally treated not as definitive refutations but as instances of wit, typically involving the placement of work by earlier authors in an uncharitably ironic light. And it would be understood as such by those who witnessed the public defense of the dissertation. The elevation of a candidate to doctoral status did not by itself constitute a change in the course of collective inquiry. One could easily imagine a future doctoral candidate ironically dismissing that doctor's arguments by reasserting the authority of the old texts. In this respect, knowledge in the world of scholarship certainly grew but without demonstrating any overall direction. Much of the humanities, including philosophy, may be seen as still operating in this mode.

Underlying the shift in mentality from scholarship to research vis-à-vis the doctoral dissertation was the significance accorded to originality, which William Clark traces to the reform of the German universities in the early 19th century. The modern requirement that dissertations constitute an "original body of research" is meant to capture two somewhat different ideas: First, there is the Romantic view—very much part of the modern German university's ideological perspective—that research should be self-motivated and self-sustaining, something done for its own sake and not to serve some other interest. But second, there is the more scholarly classical view of originality that one gets closer to the original sources of knowledge by eliminating various interpretive errors and placing salient truths in the proper light. Isaac

Newton, who by the mid-18th century was already touted as the greatest human intellect ever to have lived, was seen as original in both senses: On one hand, he laid out a conceptual framework for understanding all matter in motion based on principles that he himself had derived. On the other hand, he presented this framework as the *prisca sapientia* (pristine wisdom) implied in Biblical theology, once shorn of its accumulated corruptions and misinterpretations.

A final trace from scholarship to research is provided by *salaried employment*. Even today scholarship may conjure up images of leisure, specifically of someone with the time to spend on reading and writing what he or she pleases. There is no sense of competition with others to reach some commonly agreed goal, let alone a goal that might provide untold benefits to all humanity. Rather, scholars set their own pace along their own intellectual trajectories. In contrast, it is rare for a research agenda to be pursued by an individual working alone. Nevertheless, just because research tends to be focused on problems of common interest, it does not follow that solutions will be quickly forthcoming. Chance plays a strong role in whether research effort pays off for the people doing it. Indeed, major research breakthroughs typically come after many people have been working in the same field, proposing competing hypotheses, for some time. Even Newton famously admitted that he saw as far as he did because he stood "on the shoulders of giants." It was just this long-term perspective on the fruits of research that justified the need for researchers to be given regular salaries regardless of the ultimate significance of what they produced.

Several historical developments are indicative of this shift in mentality. The segregation of teachers and researchers in separate institutions starting in Napoleonic France enabled research to be conceptualized as a full-time job. In addition, the increasing need for complex specialized equipment in the natural sciences, starting with chemistry, forced a greater dependency of academics on support from the business community, which in turn opened the door to knowledge work being seen as a form of high-skill, high-tech industrial labor. The precedent was set by Justus Liebig's establishment of the first university-based laboratory at the Hessian University of Giessen in 1840, soon followed by Lord Kelvin at the University of Glasgow. A downstream effect, evident in the global ascendancy of German academic research in the final quarter of the 19th century, was what William Clark has identified as the Harnack principle, named after the minister for higher education who argued that research institutes should be organized around individuals with ambitious yet feasible research plans, even if others end up being employed to do most of the work. This principle is still very much in force in the funding of research teams through the vehicle of the principal investigator who acts as an entrepreneur who attracts capital investment that then opens up employment opportunities for researchers.

## Research as a Collective Project of Indefinite Duration

As a sociologist concerned with the emergence of the role of the scientist in society, Joseph Ben-David stressed the significance of the research mentality, specifically that truth about some domain of reality is worth pursuing, even if it is very unlikely to be achieved in one's lifetime, because others can then continue the project. Ben-David was struck by the cultural uniqueness of this idea of "science as a vocation," which in the 20th century was most closely associated with Max Weber. After all, many of the metaphysical ideas and empirical findings of modern science were also present in ancient Greece, India, and China, yet by our lights there seemed to be relatively little appetite for improving and adding to those ideas and findings. If anything, the normative order of these societies tended to inhibit the free pursuit of science because of its potentially destabilizing consequences. In effect, they lacked a sense of the scientist as a distinct kind of person whose primary social identity might derive from affiliations with like-minded people living in other times and places.

Indicative of the establishment of such an identity is the gradual introduction in the 19th and 20th centuries of academic journals affiliated not with particular universities or even national scientific bodies but with internationally recognized scientific disciplines. This trend greatly facilitated the conversion of scholarship to research by introducing criteria of assessment that went beyond the

local entertainment value of a piece of intellectual work. One way to envisage this transformation is as an early—and, to be sure, slow and imperfect—form of broadcasting, whereby a dispersed audience is shaped to receive the same information on a regular basis, to which audience members are meant to respond in a similarly stylized fashion. This transformation served to generate an autonomous collective consciousness of science and a clear sense of who was ahead and behind the cutting edge of research. However, there remained the question of how to maintain the intergenerational pursuit of a research trajectory, such that one can be sure that unsolved problems in what has become a clearly defined field of research are not simply forgotten but carried forward by the next generation. Enter the textbook.

National textbooks that aspired to global intellectual reach were a pedagogical innovation adumbrated by Napoleon but fostered by Bismarck. They have served as vehicles for recruiting successive generations of similarly oriented researchers in fields that Kuhn would recognize as possessing a paradigm that conducts normal science. An unnoticed consequence of the increased reliance on textbooks has been that people with diverse backgrounds and skills can regularly become integrated into a common intellectual project. The clearest indicator is the textbook's integration of textual and visual representations. Just as text is no longer solely chasing text, an original medieval practice still common in most humanistic scholarship, so too craft-based knowledge required for experimental design is no longer treated as inherently tacit and hence subject to esoteric rites of guild-style apprenticeship. Arguably the textbook has done the most to convert modern science's research mentality into a vehicle for the democratization of knowledge production.

*Steve Fuller*

***See also*** Applied Research; Ethics in the Research Process; Qualitative Research; Quantitative Research

**Further Readings**

Ben-David, J. (1971). *The scientist's role in society.* Englewood Cliffs NJ: Prentice Hall.

Ben-David, J. (1992). *Scientific growth.* Berkeley: University of California Press.

Cahan, D. (Ed.). (2003). *From natural philosophy to the sciences.* Chicago: University of Chicago Press.

Clark, W. (2006). *Academic charisma and the origins of the research university.* Chicago: University of Chicago Press.

Collins, R. (1998). *The sociology of philosophies: A global theory of intellectual change.* Cambridge MA: Harvard University Press.

De Mey, M. (1982). *The cognitive paradigm.* Dordrecht, the Netherlands: Kluwer.

Fuller, S. (1988). *Social epistemology.* Bloomington: Indiana University Press.

Fuller, S. (2000). *The governance of science.* Milton Keynes, UK: Open University Press.

Fuller, S. (2000). *Thomas Kuhn: A philosophical history for our times.* Chicago: University of Chicago Press.

Fuller, S. (2002). *Knowledge management foundations.* Woburn, MA: Butterworth-Heinemann.

Fuller, S. (2007). *The knowledge book: Key concepts in philosophy, science and culture.* Chesham, UK: Acumen.

Fuller, S. (2009). *The sociology of intellectual life.* London: Sage.

Fuller, S., & Collier, J. (2004). *Philosophy, rhetoric and the end of knowledge* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum. (Original work published 1993 by Fuller)

Grafton, A. (1997). *The footnote.* Cambridge, MA: Harvard University Press.

Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago: University of Chicago Press. (Original work published 1962)

McNeeley, I., & Wolverton, L. (2008). *Reinventing knowledge: From Alexandria to the Internet.* New York: Norton.

Price, D. de S. (1986). *Big science, little science, . . . and beyond* (2nd ed.). New York: Columbia University Press. (Original work published 1963)

Proctor, R. (1991). *Value-free science? Purity and power in modern knowledge.* Cambridge, MA: Harvard University Press.

## Research Design Principles

Research design is the plan that provides the logical structure that guides the investigator to address research problems and answer research questions. It is one of the most important components of research methodology. Research methodology not only details the type of research design to be implemented but includes

the approach to measuring variables and collecting data from participants, devising a strategy to sample participants (units) to be studied, and planning how the data will be analyzed. These methodological decisions are informed and guided by the type of research design selected. There are two broad categories of research design: observational and interventional. Each research design emphasizes a different type of relationship to be explored between variables. The social issue or problem that needs further exploration and study is articulated in the research problem statement, which conceptualizes the problem or issue and identifies the main salient variables to be examined. Based on this conceptualization, research questions are developed to isolate the specific relationships among variables to be assessed: correlational and causal. The kind of research questions that the investigator wishes to answer will help determine the specific type research design that will be used in the scientific inquiry. Determining what kind of research questions should be posed and answered is one of the most important decisions to be made in the planning stages of a research study. It provides the foundation for selecting the actual design to be used in the study.

## Research Problem

The research problem is an area of concern that the investigator is interested in exploring. It reflects a gap in knowledge about a specific social issue or problem discussed in the scientific literature. Often a research problem focuses on a situation that arises out of everyday life and that calls for a solution, an improvement, or some alteration; it is a problem that needs to be solved. Conceptualizing and describing the social problem lead the investigator to identify the most salient concepts or variables involved and guide the investigator's thinking about why and how the variables are connected to each other. Through this process of articulating and framing the research problem, investigators may discover that little is known about an issue or that there is conflicting information about it within the scientific literature. Depending on what is known about the research problem, the types of specific research questions to be asked and the purpose and objective of the research study could range from basic exploratory research to control experiments.

## Research Questions and Hypotheses

Research questions involve examining a relationship between two or more variables. The relationship between variables can be expressed as either positive or negative correlation or causal. An example of a research question about a correlation would be, Are neighborhood characteristics of crime and low-income related to health status?

Causal relationships describe the cause-and-effect relationships between variables. Three criteria are used to assess causal relationships: (1) correlation between the variables, (2) temporal precedence or sequence, and (3) no alternative explanation. First, there must be a significant positive or negative correlation between the two variables. That is, as the value of one variable changes, the value of the other changes as well. Second, there is a defined temporal sequence of the variables. The variable that is considered to be the cause (the independent variable) must precede the effect variable (the dependent variable) in time. The temporal relationship is often diagrammed as $A \rightarrow B$ or expressed as, if A then B. Third, the cause-and-effect relationship between the variables cannot be plausibly explained by the presence of another variable. This is called the *third variable problem,* which is reduced or eliminated by controlling specific methodological procedures used to implement a particular research design. These controls attempt to combat specific threats to internal validity that can negatively affect the inferences made from a study. Examples of research questions examining a causal relationship are, Does cognitive behavioral therapy reduce trauma symptoms among survivors of domestic violence? and Does neighborhood cohesion reduce community crime?

When there is an existing theory and/or the empirical literature provides some evidence about the relationship between variables, then an investigator is able to venture a hypothesis. A hypothesis is a formal statement about the expected relationships between two or more variables in a specified population. It is a translation of a research problem into a clear explanation or prediction of a study's results. A hypothesis is stated in a manner that allows for it to be tested, that is, either

supported or refuted. Examples of hypotheses are; Cognitive behavioral therapy will reduce trauma symptoms among survivors of domestic violence, and, The higher the neighborhood cohesion, the more community crime will be reduced.

Thus, research questions and hypotheses examine whether variables stated in a research problem are related or have a causal relationship. The types of questions and hypotheses that are expressed are strongly linked to the purpose of the research study.

## Purpose of Research

Research can be grouped into exploratory, descriptive, explanatory, and evaluative. Research that is exploratory or descriptive examines the relationships among variables (correlatonal), and research that focuses on explanation or evaluation assesses causal relationships.

*Exploratory studies* are usually undertaken when relatively little is known about a phenomenon or an event. The goal is to determine whether a relationship exists among several variables under scrutiny. This type of study often operates as a pilot study to determine the feasibility of conducting a larger, more complex study. For example, investigators may be interested in exploring whether there is a relationship between neighborhood conditions and a person's physical health. Without making a simple connection between neighborhood conditions and health, implementing an intervention study to change the conditions would be premature, because it would be unknown which neighborhood condition to focus on (e.g., economic, political, social, psychological, housing, community resources, crime).

*Descriptive research* attempts to describe a group of people, a phenomenon, or an event. It is one of the first steps in understanding social problems and issues; it describes who is experiencing the problem, how widespread the problem is, and how long the problem has existed. An example of a simple descriptive study is an investigation of how many people within a community are men, have less than a high school education, or earn less than the poverty threshold. Another purpose is to describe the prevalence of a phenomenon or event. These research questions might ask, What is the prevalence of methicillin-resistant *Staphylococcus aureus* (a type of bacterial infection that is resistant to antibiotics and difficult to treat) in the community? or What is the prevalence of homelessness in New York City?

*Explanatory research* focuses on why variables are related to one another. One main purpose of explanatory research is to test theory. A theory can be defined as a set of logically organized and interrelated propositions, statements, principles, rules, and assumptions of some phenomenon that can be used to describe, explain, predict, and/or control that phenomenon. Many theories describe essential cause-and-effect relationships between variables. They posit the direction (positive or negative), the strength, and the temporal (causal) sequence between variables. In an explanatory research study, the investigator measures these variables and then provides evidence that either supports or refutes the contention that there is a cause-and-effect relationship that exists between these variables. Specific research designs, such as experimental and prospective (longitudinal), provide an opportunity to observe and assess causal relationships directly. It is important to perform descriptive or exploratory research that describes the prevalence of a specific social problem, such as crime, substance abuse, or homelessness, and establishes basic relationships between variables before conducting more expensive and time-consuming research. Explanatory research can lead to research questions about how these conditions can be changed.

*Evaluative research* attempts to assess whether a particular intervention, process, or procedure is able to change behavior. Program evaluation falls under this type of research. Its goal is to evaluate the effectiveness of social programs that provide services. Program evaluations try to establish whether the intervention actually changed people's behavior and is responsible for improvements in client outcomes. For example, a federal educational program that has that existed in the United States since 1965 is Head Start. An evaluation of Head Start could determine whether providing comprehensive education, health, nutrition, and parent involvement services to low-income children and their families can improve important developmental and education outcomes for preschool children.

Once the problem has been defined, the questions articulated and the purpose of the research set, the next decision is to choose a research design that reflects these factors.

## Research Designs

Research designs fall into two broad categories: observational and interventional. These designs vary in their approach to answering research questions and in how they depict relationships among variables. Observational designs examine the association among variables, and interventional designs focus on cause-and-effect relationships between variables. Thus, once an investigator poses a research question, then the choice of a specific research design becomes clearer.

### *Observational*

Observational designs are referred to as *nonexperimental* because the investigator does not intervene or manipulate any variables. The most common observational designs are cross-sectional, cohort, and case–control, in which data can be collected at one time, prospectively (longitudinally), or retrospectively.

#### Cross-Sectional

In cross-sectional designs, the investigator collects all the data at the same time. These designs have been used extensively to describe patterns of variables within a population. Annually, U.S. federal agencies such as the Census Bureau, Centers for Disease Control and Prevention (National Center for Health Statistics), and Bureau of Labor Statistics conduct several nationally representative cross-sectional studies to describe and monitor the physical health, mental health, substance use, morbidity and morality, births, health insurance, housing, employment, income, and other key characteristics of the population. In addition, survey organizations such as the National Opinion Research Center (NORC) and the Pew Research Center use cross-sectional designs to survey the U.S. population in order to monitor key social, economic, political, and health indicators. Many of these surveys are available for secondary data analysis.

#### Cohort Design

The word *cohort* is derived from *cohor,* which was 1 of 10 divisions making up a Roman legion. In research, the term refers to any group of individuals with something in common. That is, a collection or sampling of individuals who share a common characteristic, such as the same age, sex, or neighborhood. Epidemiological and clinical research identify a cohort as a well-defined group of people or patients who share a common experience or exposure and who are then followed for a time to examine the incidence of new diseases or events. Cohort designs are typically conducted either prospectively or retrospectively.

*Prospective Cohort Design.* In prospective cohort design, a cohort of individuals is followed over an extended time, such as 1 year, 5 years, or 10 years. The potential risk factors (predictors) and outcomes are collected from each participant during this time at a regularly scheduled interval, such as annually or every 2 years. Collecting data using a structured schedule allows the investigator to observe whether changes in potential risk factors precede changes in specific outcomes. By establishing a temporal sequence, the investigator is able to provide some evidence to determine whether the relationship between predictors and outcomes is potentially causal. The Framingham Heart Study is an example of a prospective study. The study began following a cohort of 5,209 men and women (between the ages of 30 and 62) who lived in Framingham, Massachusetts, in 1948. The purpose of the study was to identify a set of common risk factors that contributed to the development of cardiovascular disease. At the beginning of the study, individuals who did not have a history of a heart attack or stroke or any overt symptoms of cardiovascular disease were given a baseline physical examination and a lifestyle questionnaire. This cohort was then followed every 2 years with a detailed medical history, physical examination, and laboratory tests. The Framingham Heart Study has provided substantial evidence for the relationship between various risk factors and the development of heart disease

*Retrospective Cohort Design* A retrospective cohort design uses data that were collected in the

past to determine the relationship between a set of potential risk factors and an outcome that is measured either in the past or the present. An example would be a study to examine the health of a community that was exposed many years ago to an industrial accident that released a harmful substance such as radiation (as in the Chernobyl disaster in Russia), a toxic gas cloud (such as the Bhopal disaster in India), or toxic waste (as in the Love Canal disaster in New York). The study would use past records (e.g., medical, school, employment, and insurance, and level of exposure to the toxin), questionnaires, or both to assemble a set of past risk factors that could be related to a person's current health status.

#### Case–Control Design

Case–control studies compare potential risk factors of individuals who already have a specific condition or disease (*cases*) with individuals who do not (*controls*). The selection of study subjects is based on the outcome variable (condition or disease). Controls are selected and matched to the cases by important demographic variables (age, gender, ethnicity, socioeconomic status). The goal of the matching is to make the controls as comparable to the cases as possible on the most important characteristics. The controls serve as a reference group with which to compare the cases on potential risks factors. Differences between cases and controls on these risk factors could suggest factors that either elevate or reduce the risk of the occurrence of a disease or condition. Thus, the investigation goes from the effect (case status) to the potential cause (risk factors). An example of a case–control study would be a comparison of children with (cases) and without (controls) attention deficit/hyperactivity disorder (ADHD) and their prenatal exposure to maternal alcohol, smoking, and drug use during pregnancy. If children with ADHD have a higher prevalence of these risk factors than do comparable children without ADHD, then there may be a potential causal connection between the exposure and ADHD.

### *Interventional Designs*

Interventional designs focus directly on an examination of the cause-and-effect relationship between independent and dependent variables. Their purpose is to create a situation in which the investigator is able to observe the change in a dependent variable (outcome) as the result of the introduction or manipulation of a stimulus, called the *independent variable.* Interventional designs are able to control or reduce the influence of other, confounding variables (called threats to internal validity) that may impact the relationship between independent and dependent variable. Interventional designs can be categorized as experiments, quasi-experiments, preexperiments, or natural experiments.

#### Experimental Designs

Experiments are designed to assess the causal relationship between an independent and a dependent variable. The classical experimental design is the *pre–post test control group,* in which individuals are randomly assigned to an experimental or control group and then compared on an outcome variable. Experimental designs assist the investigator in reducing or controlling many sources of bias and threats to internal validity that may influence the inference made about a cause-and-effect relationship between the independent and dependent variable. A defining feature of these experiments is *random assignment* of participants to experimental or control groups. The goals of random assignment are to ensure that subjects have an equal chance (or at least known probability) of being selected for either an experimental or a control group and that these groups are as similar as possible to each other in personal characteristics and outcome variables prior to any intervention. The latter criterion is called the *equivalence of groups.* Random assignment lets chance place individuals into groups rather than allowing an investigator to choose who should be in a specific group. This approach does not guarantee group equivalence, but it does reduce selection bias.

After random assignment, the experimental group would be exposed to the intervention or stimulus, but the control group would just be measured over the same time. If the experimental group significantly changes on the dependent variable from the baseline measurement to the posttest measurement but the control group remains relatively the same, then it could be concluded or inferred that the intervention or the stimulus may be causing

the difference between the two groups. In order for this inference to be considered, all other sources of bias or internal validity threats must be controlled for during the experiment. If the two groups do not differ from one another on the dependent variable, then it can be concluded that either the intervention did not work or that some form of error or bias was introduced by the study's procedures and impacted the validity of the study.

Educational, social, and psychological experiments introduce a stimulus to examine its influence on animal or human behavior. For example, an experiment could assess the effect of noise (stimulus) and the performance of a task (dependent variable), such as number of recalled new words from a list. For various psychosocial problems or conditions, experimental designs are used to assess existing and new treatments that are purported to improve an individuals' behavior or quality of life. For example, an experiment could be designed to examine a new approach for treating people with hypochondriasis. The experimental group would receive cognitive-behavioral therapy, and the control group would receive usual care. By comparing groups on several outcome variables, such as the level of health anxiety or number of visits made to primary care physicians, it could be shown that the treatment approach was similar, better, or worse than the usual care in helping individuals with hypochondriasis.

There are many types of experimental designs, beyond the classic pre–post test control group, that are used to control or minimize various biases and threats to ensure that the inference made about the cause-and-effect relationships from a study are valid. These designs include the following: posttest only control group, alternate treatments, multiple treatments and controls, Solomon four group, Latin square, crossover, factorial, block, and repeated measures (longitudinal, nested (hierarchical), mixed).

### Quasi-Experimental Designs

Similar to experimental designs, quasi-experiments manipulate an independent variable and compare the performance of at least two groups on an outcome variable. However, quasi-experimental designs do not have some of the elements of control that are found in experiments, and thus the groups are referred to as intervention and comparison. The main difference between quasi-experimental and experimental designs is the lack random assignment. Not randomly assigning participants to groups can result in the nonequivalence of groups. That is, the intervention and comparison groups may begin the study with differences in personal characteristics and/or the outcome variable. These differences can reduce the ability to make inferences about causal relationships. Quasi-experimental designs are used when it is not feasible to perform a true experiment. Often random assignment is not possible for ethical reasons or because of policies of social programs that will not allow interventions to use random assignment, whereby some clients would not receive any form of treatment (viewed as withholding effective treatment). The investigator would have to use statistical techniques such as analysis of covariance or propensity scoring to statistically reduce the bias due to these covariates that lead to initial group differences. An example of a quasi-experimental study would be a comparison of the effectiveness of two treatment approaches, Housing First and Supportive Housing, to help individuals who are homeless and have both a mental illness and a substance abuse disorder. Two emergency shelters, whose administrators are opposed to random assignment, could be used as sites in the study. Site A could be provided with the Housing First intervention and Site B could be given the Supportive Housing intervention. The two interventions would be compared at the end of the study on outcome measures such as housing stability and duration, quality of life, and use of psychiatric hospitalizations. However, lacking random assignment, these sites could vary significantly in many client characteristics before the study even begins, thus affecting any inferences of causality regarding the intervention.

### Pre-Experimental Designs

Pre-experimental designs are designs that do not share the same strong features of experimental or quasi-experimental designs that control for internal validity threats. These designs typically lack a baseline measurement of important outcomes and/or a comparison group. The *one-shot case study* examines a group of participants after the introduction of an intervention or stimulus;

however, there is no ability to compare to baseline measures or other groups to establish a simple correlation. The *one-group pretest–posttest design* has pre–post measurements of the important independent and dependent variables; however, it lacks a control group for comparison. The *static-group comparison design* has a posttest measurement of the dependent variable on the one group and compares this group to a nonequivalent group that did not receive the intervention. However, the lack of random assignment and/or baseline measurements of the dependent variable limit the inferences that may be made. These designs are considered the weakest designs that try to establish causal relationships between independent and dependent variables.

#### Natural Experiment

Natural experiments are not true experiments, because the investigator does not manipulate the independent variable. Events such as natural disasters, wars, environmental accidents, or manufacturing plant closings that occur in everyday life can dramatically change human behavior. All these events can have a devastating impact on individuals, families, and communities. For example, hurricanes and floods can change a community with great ferocity and speed by destroying homes, businesses, and communities, as well as taking the lives of family, friends, and neighbors. Hurricane Katrina, in August 2005, was one such event that affected Gulf coast states including Louisiana, Mississippi, Alabama, and Florida, as well as the city of New Orleans. A natural experiment would involve studying the impact of the hurricane on individuals' psychological, social, and physical health by comparing a community that was directly impacted by the hurricane with a similar community that was not.

## Threats to Validity

Validity focuses on the approximate truth of an inference. In a study, many inferences can be made about what caused the relationship among the variables. However, the investigator is never certain which inferences are true and which ones are false. By controlling the many threats to validity, the investigator is able to eliminate many variables that can influence the results of a study and lead to false inferences that can distort the relationship between independent and dependent variables. These threats are often referred to as *alternative explanations*.

There are four main sources of threats: internal, statistical conclusion, construct, and external. *Internal validity* focuses on what occurred during the implementation of the study that could influence the relationship between the independent and dependent variables. Threats to internal validity involve specific design elements that impact the temporal precedence of the independent and dependent variables and the selection of participants, as well as the effects of history, maturation, differential attrition, testing, and instrumentation on the outcomes of the study. *Statistical conclusion validity* involves inferences about the correlation or covariation between an independent and dependent variable. The power of a study is often examined as one of the main issues: does study have a sufficient number of subjects to detect a relationship if one exists? *Construct validity* focuses on the inferences made about the higher ordered constructs represented in the study by the independent and dependent variables. *External validity* refers to the generalizability of the findings to other persons, settings, or times.

The choices an investigator makes of the type of design to be used and how the study will be implemented will impact the overall validity of the study. These decisions are often trade-offs between theoretical, practical, and ethical issues in conducting a study. Further, correcting or adjusting for one threat may increase the likelihood of another threat.

*Bruce R. DeForge*

***See also*** Block Design; Cohort Design; Crossover Design; Cross-Sectional Design; Experimental Design; Factorial Design; Latin Square Design; Quasi-Experimental Design; Repeated Measures Design; Threats to Validity

#### Further Readings

Bernard, H. R. (2000). *Social research methods: Qualitative and quantitative approaches*. Thousand Oaks, CA: Sage.

Black, T. R. (2003). *Doing quantitative research in the social sciences: An integrated approach to research*

*design, measurement and statistics*. Thousand Oaks, CA: Sage.
deVaus, D. (2001). *Research design in social research*. Thousand Oaks, CA: Sage.
Engel, R. J., & Schutt, R. K. (2005). *The practice of research in social work*. Thousand Oaks, CA: Sage.
Hulley, S. B., Cummings, S. R., Browner, W. S., Grady, D., Hearst, N., & Newman, N. (2007). *Designing clinical research: An epidemiological approach* (3rd ed.). Philadelphia: Lippincott, Williams & Wilkins.
Kerlinger, F. N., & Lee, H. B. (2000). *Foundations of behavioral research* (4th ed.). Fort Worth, TX: Harcourt College.
Neuman, W. L., & Kreuger, L. W. (2003). *Social work research methods: Qualitative and quantitative applications*. Boston: Allyn & Bacon.
Rubin, A., & Babbie, E. (2007). *Research methods for social work* (6th ed.). Belmont, CA: Wadsworth.
Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimentation designs for generalized causal inference*. Boston: Houghton Mifflin.
Trochim, W. K., & Donnelly, J. P. (2007). *The research methods knowledge base* (3rd ed.). Mason, OH: Thomson.

## Research Hypothesis

To conduct research is to collect, analyze, and interpret data systematically so as to answer specific questions about a phenomenon of interest. These questions may be derived from conjectures about (a) an efficacious cause that brings about the phenomenon, (b) the intrinsic nature of the phenomenon, or (c) how the phenomenon is related to other phenomena. Tentative answers to these questions are research hypotheses if they are (a) consistent with the to-be-explained phenomenon, (b) specific enough to serve as guidelines for conducting research, and (c) testable (i.e., there are well-defined criteria of rejection).

Depending on the level of abstraction adopted, underlying an empirical research are three hypotheses at different levels of theoretical sophistication or specificity, namely, *substantive, program,* and *individual* research hypotheses (see Table 1).

A practical problem or a new or intriguing phenomenon invites speculations about its cause or nature or relation to other phenomena. For example, it is commonly accepted that people in small townships are friendlier than their counterparts in bigger cities. To investigate whether or not this is the case, as well as the reason, researchers would first offer a speculation (e.g., environmental effects) that, if substantiated empirically, would explain the phenomenon (see row 1 of Table 1). Such a conjecture is a *substantive hypothesis* because it explains a real-life (substantive) phenomenon (see row 2).

Substantive hypotheses are typically too general or vague to give directions to empirical research. Part of the reason is that the phenomenon is multifaceted. For example, environmental factors are air quality, noise level, amenities of various sorts, traffic volume, and the like. By itself, any one of these environmental factors also has multiple components. Hence, an investigation of a substantive hypothesis implicates a program of related hypotheses. Such a related set of hypotheses may be characterized as *program hypotheses* (see $R_1$, $R_2$, and $R_3$ in row 3).

To the extent a program hypothesis is well defined, researchers can conduct an experiment by specifying the to-be-used research method, materials, procedure, and to-be-measured behavior (see [i] in row 4a). Specifically, the independent variable envisaged is *Ion M,* whose two levels are *presence* and *absence.* A hypothesis of such specificity is an individual research hypothesis, which may be in the form of an experimental hypothesis (see $E_1$ in row 4b).

*Ion M* is the independent variable in the present example (a) because Implication $R_1$ (namely, $E_1$) in [i] of row 3 is adopted, and (b) because of the methodological assumption that *Ion M* cleanses the air. Had $R_2$ in row 3 been used instead, a different independent variable (e.g., food supplement *H*) would be used (e.g., see [ii] in rows 4a and 4b).

In short, *research hypothesis* may refer to any one of three hypotheses underlying an empirical research study, namely, the substantive, program, and individual study hypotheses. Although the three hypotheses are literally different and may ostensibly be about different things, they are theoretically bounded. Specifically, the substantive hypothesis implies any one of the program hypotheses, which, in turn, implies one or more individual study hypotheses. For these implicative relations to be possible, the substantive or research hypothesis must be well defined. An individual *study hypothesis* of sufficient specificity becomes an experimental hypothesis. At the same time,

**Table 1** The Substantive, Program, Research, Experimental, and Statistical Hypotheses

| Level of Abstraction | | Phenomenon or Question of Interest | Conjecture or Hypothesis |
|---|---|---|---|
| 1 | Phenomenon | Residents in small cities are friendlier than their big-city counterparts. | Environments in big cities are less pleasant than in small towns. |
| 2 | Substantive | How would a more pleasant environment make people friendlier? | $T_1$ : Environmental pleasantness eliminates irritability. |
| 3 | Program | Would less irritable people be more polite?<br>Would cultural enrichment make people more polite?<br>Would less traffic make people more polite? | $R_1$ : Less irritated people are more polite.<br>$R_2$ : Antibiotics in meat render people irritable.<br>$R_3$ : People are more polite in situations of lower traffic density. |
| 4a | Research manipulation | [i] Suppose that *Ion M* eliminates air pollution. *Ion M* is used in the experimental condition (the *Ion M Present* condition), whereas there is no *Ion M* in the control condition (the *Ion M Absent* condition).<br>[ii] Suppose that *Supplement H* in food eliminates antibiotics in the body. *Supplement H* is used in the experimental condition (the *Supplement H Present* condition), whereas there is no *Supplement H* in the control condition (the *Supplement H Absent* condition). | |
| 4b | Individual study (e.g., an experiment) | [i] Would people be more polite in a condition with cleaner air?<br>[ii] Would people be more polite when they have fewer antibiotics in their system? | $E_1$ : People use more polite words in a condition with cleaner air.<br>$E_2$ : People use more polite words when they have fewer antibiotics in their system. |
| 5a | Statistical | Are the data the result of chance influences? | $H_0 : \mu_E \leq \mu_C$<br>$H_0$ : The mean number of polite words used in the *Ion M Present* condition is smaller than or equal to that in *Ion M Absent* condition. |
| 5b | | Are the data the result of non-chance influences? | $H_1 : \mu_E > \mu_C$<br>$H_1$ : The mean number of polite words used in the *Ion M Present* condition is larger than that used in *Ion M Absent* condition. |

*Notes:* $E$ = individual research hypothesis; $R$ = program hypothesis; $T$ = substantive hypothesis; Subscript $E$ = *Ion M Present;* Subscript $C$ = *Ion M Absent*.

even when one is collecting data to test a specific experimental hypothesis, one's ultimate conclusion is about the substantive hypothesis via the research hypothesis.

Note that the experimental hypothesis, let alone the research or the substantive hypothesis, is neither the statistical alternative ($H_1$; see row 5b) nor the statistical null hypothesis ($H_0$; see row 5a). While the nonstatistical hypotheses are conceptual conjectures about a phenomenon or the conceptual relationships between the independent and dependent variables, the statistical hypotheses are hypotheses about the effects of chance influences on research data. That is, the conceptual and statistical hypotheses implicated in a research belong to different domains.

*Siu L. Chow*

***See also*** Experimental Design; Replication; Rosenthal Effect

# Research Question

One of the most important decisions a researcher will make is deciding on a question that will be examined in the research process. This can be a very difficult decision because there can be many questions a researcher wants and needs to address. Good supervision will guide a researcher to a focused area of study that can lead to a focal question. The debate continues on whether this focus should be singular or plural, and the decision can be influenced by the subject of study and various quantitative or qualitative methodologies employed. John Creswell has suggested that before searching the literature, a researcher should identify a topic, using such strategies as drafting a brief title or stating a central research question. Researchers need to be aware of the following issues in relation to creating the research question.

## Choosing the Right Research Question

Choosing the right research question involves a number of decisions that will shape both the research project and process. The "right" question has to involve several issues: an adequate knowledge of the area being considered for research, constructive support from a supervisor, and the time a researcher has to carry out the research. First, the researcher must have knowledge of the general and specific subject area. This knowledge may have been obtained from previous study; otherwise, research evidence from recent, published literature has to be examined. The researcher needs to realize that the research to be undertaken contributes to ongoing debates within a subject area being researched. Some students and academics do not realize that the rules and regulations that apply to senior researchers also apply to students and novice researchers at the undergraduate level (termed *junior researchers*). Literature reviews have to begin at the earliest possible stage of research because this knowledge has to inform the development of the research question. The first evidence base in research is published and edited literature, and the application of key and core texts within the subject has to inform the creation and evolution of the research question. This strategy suggests that the research question might have to change when a researcher acknowledges what has been published in a subject area.

A supervisor plays a crucial role in directing or nudging the researcher in the right direction. As Thomas Murray and Dale Brubaker suggest, it is well to identify potential sources of help and to recognize the advantages and limitations of each. Those sources of most value are usually academic advisers, fellow graduate students, experts outside one's own department or institution, the researcher him- or herself, and the professional literature. The supervisor can advise which research direction a researcher can travel. This direction can be absolutely controlled if the researcher is working within a funded team, in which a supervisor can be a senior researcher. Relative control exists when a research supervisor informs the researcher about research design, which consists not only of devising the right question but also of literature review; methods and methodology, data collection, presentation, analysis, and answering the question in a research project conclusion. Third, time is an issue that has to be considered by all researchers when thinking how to answer a research question. On reflection, planning and organization are crucial techniques in research design, but time still shapes what can and what cannot be done. As Elizabeth Wilson and Dorothy Bedford have argued, What are the consequences for full- and part-time students and staff who have full-time responsibilities outside academia? Time is an issue that also concerns both the quality of research and supervision.

## Justifying the Research Question

Following these three issues, the researcher must justify which subject and question or hypothesis is going to be researched. Justification involves choice. Which decisions have to be made before a research project can begin? These choices involve how specific a subject area needs to become when addressing a problem. Which problem can be addressed when considering knowledge, supervision, and time? That narrowing down or becoming focused on a research area is possibly one of the most important processes involved within research because justifying the research project makes it not only viable but indeed possible. Focusing on a research area allows the next focal decision, which involves which question can be

asked and addressed. Even at this early stage, ethical issues need to be considered. For example, what happens when a research project involves participants age 16 or younger in a methodology? Which questions can be asked? What experiments can take place within a laboratory?

The type of research question needs to also be addressed. Patrick White divides question types into descriptive and explanatory, the W-questions (who, what, where, when, why, and how), and purpose-led typologies. The purpose issue also helps in finding that focus and reducing the number of questions or subquestions. This is subject oriented, and different subjects and disciplines have different methods in relation to these research choices. Justification of research question is both subject based and situational. For example, a dissertation needs to be focused within roughly a 6-month time frame. If time is longer, then the purpose can be multifocal, but a single research question may produce more focused results and recommendations for further research.

*Richard Race*

*See also* Literature Review; Proposal; Research; Research Design Principles; Research Hypothesis

**Further Readings**

Creswell, J. W. (2009). *Research design: Qualitative, quantitative, and mixed methods approaches* (3rd ed.). Thousand Oaks, CA: Sage.

Murray T. R., & Brubaker, D. L. (2008). *Theses and dissertations* (2nd ed.). Thousand Oaks, CA: Sage.

Race, R. (2008). Education, qualitative research in. In L. M. Given (Ed.), *Encyclopedia of qualitative research methods* (pp. 240–244). Thousand Oaks, CA: Sage.

White, P. (2009). *Developing research questions: A guide for social scientists*. New York: Palgrave Macmillan.

Wilson, E., & Bedford, D. (2009). *Study skills for part-time students*. Harlow, UK: Pearson Education.

# Residual Plot

Residual plots play an important role in regression analysis when the goal is to confirm or negate the individual regression assumptions, identify outliers, and/or assess the adequacy of the fitted model. Residual plots are graphical representations of the residuals, usually in the form of two-dimensional graphs. In other words, residual plots attempt to show relationships between the residuals and either the explanatory variables ($X_1$, $X_2, \ldots, X_p$), the fitted values ($\hat{y}_i$), index numbers $(1, 2, \ldots, n)$, or the normal scores (values from a random sample from the standard normal distribution), among others, often using scatterplots. Before this entry discusses the types of individual residual plots in greater detail, it reviews the concept of residuals in regression analysis, including different types of residuals, emphasizing the most important standard regression assumptions.

## Concept

Defining a linear regression model as

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon, \qquad (1)$$

the $i$th fitted value, the value lying on the hyperplane, can be calculated as

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip} = 1, 2, \ldots, n, \qquad (2)$$

where $p$ refers to the number of explanatory variables and $n$ the number of observations. For $p = 1$, the regression model represents a line; for $p = 2$, a plane; and for $p = 3$ or more, a hyperplane. From here on, the term *hyperplane* is used, but the same concepts apply for a line, that is, for the simple regression model, or a plane.

The $i$th residual ($e_i$), is then defined as

$$e_i = y_i - \hat{y}_i, \qquad (3)$$

and it represents the vertical distance between the observed value ($y_i$) and the corresponding fitted values ($\hat{y}_i$) for the $i$th observation. In this sense, the $i$th residual ($e_i$) is an estimate for the unobservable error term ($\varepsilon$) and, thus, represents the part of the dependent variable ($Y$) that is not linearly related to the explanatory variables ($X_1$, $X_2, \ldots,$ $X_p$). One can also say that the $i$th residual ($e_i$) is the part of $Y$ that cannot be explained by the estimated regression model.

In regression analysis, the error terms incorporate a set of important and far-reaching assumptions. These are that the error terms are independently and identically distributed normal random variables, each with a mean of zero and

a common variance. This assumption can be expressed as

$$\varepsilon \sim NID\left(0, \sigma^2 I_n\right). \quad (4)$$

The residuals, however, do not necessarily exhibit these error term assumptions. By definition, for instance, the sum of all residuals must equal to zero, or

$$\sum e_i = 0. \quad (5)$$

The direct implication from this residual property is that although the error terms are independent, the residuals are not. In addition, residuals resemble only estimates of the true and unobservable error terms and as such do not have the same variance $\sigma^2$, as implied in Equation 4. The variance of the $i$th residual is defined as

$$Var(e_i) = \sigma^2(1 - p_{ii}), \quad (6)$$

where $p_{ii}$ is the $i$th leverage value, that is, the $i$th diagonal element in the projection matrix, also known as the hat matrix. This unequal variance property can be coped with by means of standardization of the residuals.

Standardization of the residuals is achieved by dividing the $i$th residual by its standard deviation,

$$z_i = \frac{e_i}{\sigma\sqrt{1 - p_{ii}}}, \quad (7)$$

where $z_i$ denotes the $i$th standardized residual, which now by definition has a mean of zero and a standard deviation of one. For practical purposes, however, $\sigma$ is not known and must therefore be replaced by the unbiased standard error of the estimate ($\hat{\sigma}$), where $\hat{\sigma}$ is defined as

$$\hat{\sigma} = \sqrt{\frac{\sum e_i^2}{n - p - 1}}. \quad (8)$$

Using the standard error of the estimate ($\hat{\sigma}$) for the standardization of the $i$th residual then gives us the $i$th internally studentized residual $r_i$:

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - p_{ii}}}. \quad (9)$$

Internally studentized residuals have a standard deviation of 1, that is, they all have the same variance, but they no longer result in a sum of zero. It should be further noted that when one refers to standardized residuals, one is often referring to internally studentized residuals. Given the fact that the true standard deviation ($\sigma$) is not known, and is usually replaced with an unbiased estimate, that is, the standard error of the estimate ($\hat{\sigma}$), the remainder of this discussion on residual plots also refers to internally studentized residuals.

Residual plots can be very powerful explanatory diagnostic tools when the task is to check the standard regression assumptions. The standard regression assumptions include the following:

1. the linearity assumption that the regression model is linear in its parameters $\beta_0, \beta_1, \ldots, \beta_p$;
2. error term ($\varepsilon$) assumptions, which state that the error terms are independently and identically distributed (iid) normal random variables, each with a mean of zero and a common variance:

$$\varepsilon \sim NID\left(0, \sigma^2 I_n\right) \quad \text{(see Equation 4)}$$

3. explanatory variable ($X$) assumptions, which indicate that the $X$s are supposed to be nonrandom variables, that they are measured without error, and that they are to be linearly independent of each other.

Residual plots provide valuable diagnostic insights when one is evaluating these standard regression assumptions, particularly the assumption about the error terms. Residual plots are done after a model has been fitted to available data, often in conjunction with formal statistical tests. To do so, the internally studentized residuals need to be saved first, a standard option usually available in statistical software packages. In a second step, using the saved studentized residuals allows derivation of different residual plots in order to visually explore whether violations of the standard regression assumptions may be present.

## Commonly Used Residual Plots in Regression Analysis

For the following discussion of the various types of residual plots, internally studentized residuals ($r_i$) are meant unless otherwise stated.

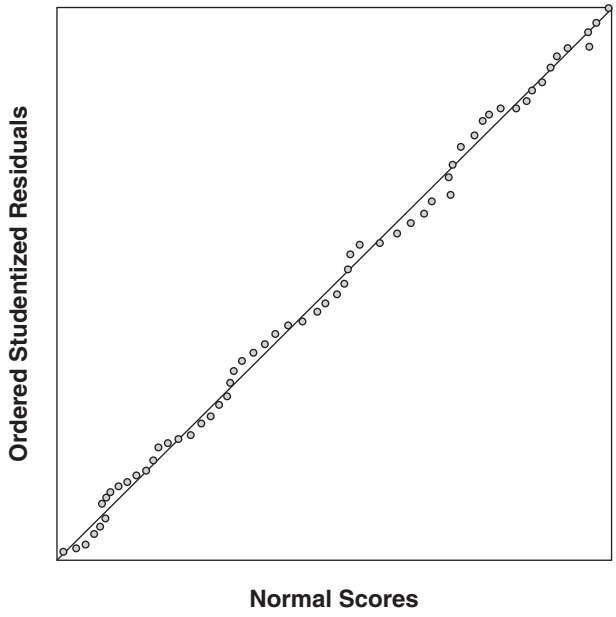



**Figure 1** Normal Probability Plot

### *Normal Probability Plot of Standardized Residuals*

To check the normality assumption, one plots the ordered residuals against the ordered normal scores. The normal scores are what one would get when taking a random sample of size $n$ from a standard normal distribution with mean 0 and standard deviation of 1. Under the normality assumption, a plot of the ordered residuals against the ordered normal scores should resemble a straight line with a slope of 1 that passes through the origin (see Figure 1).

The more the graph deviates from a straight line, the more the normality assumption is violated, and the less likely it is that the residuals are normally distributed.

### *Scatterplots of Studentized Residuals Versus the Explanatory Variables*

A very common practice in regression analysis is plotting the residuals against each of the explanatory variables ($X_1, X_2, \ldots, X_p$), meaning that for $p$ explanatory variables, there will be $p$ individual scatterplots. In regression analysis, changes in the dependent variable ($Y$) are explained through a set of explanatory variables. More specifically, it is the part in $X_1, X_2, \ldots, X_p$ that linearly relates to $Y$ that explains changes in the dependent variable ($Y$). Analogously, the part of the dependent variable ($Y$) that is not linearly related to the explanatory variables ($X_1, X_2, \ldots, X_p$) is captured by the residuals. In other words, the scatterplot of the residuals versus one of the explanatory variables exhibits a random pattern if the linearity assumption holds and there indeed exists a linear relationship between $Y$ and $X_j$, as shown in Figure 2. For those explanatory variables ($X_j$) for which there is no significant linear relationship with the dependent variable ($Y$), the scatterplot will not exhibit a random pattern. Obviously, the nonlinear relationship outweighs the linear component of the relationship of $Y$ and $X_j$, which shows in the form of a distinct nonlinear pattern, as indicated in Figure 3.

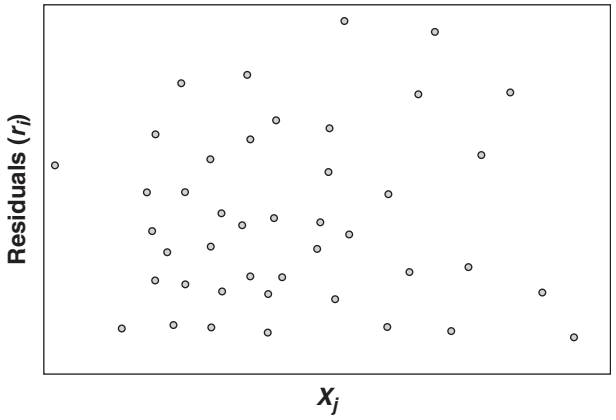



**Figure 2** Scatterplot of Residuals Versus $X$ Showing Random Pattern

One uses the same scatterplots of the residuals ($r_i$) versus the explanatory variables ($X_j$) when checking for the homogeneity of the residual variances, that is, the common variance assumptions, also referred to as *homoscedasticity*. Contrarily, *heteroscedasticity* refers to unequal error variance, that is, error variance that is not constant across all observations. The idea is that standard deviations of the error terms are constant and independent of the values of $X$. If heteroscedasticity is present, the estimated regression coefficients remain still unbiased, but their standard errors are likely to be underestimated. As a consequence, the corresponding

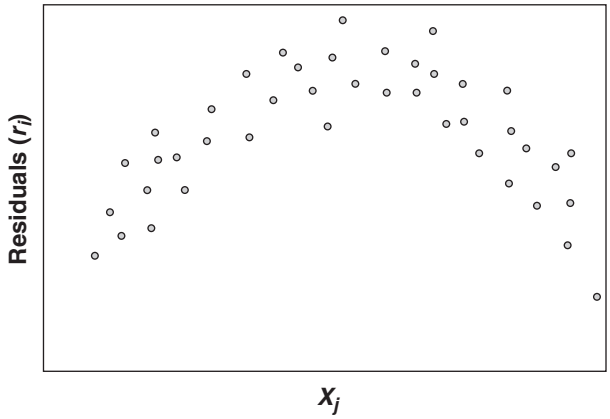



**Figure 3** Scatterplot of Residuals Versus *X* Indicating Nonlinear Relationship

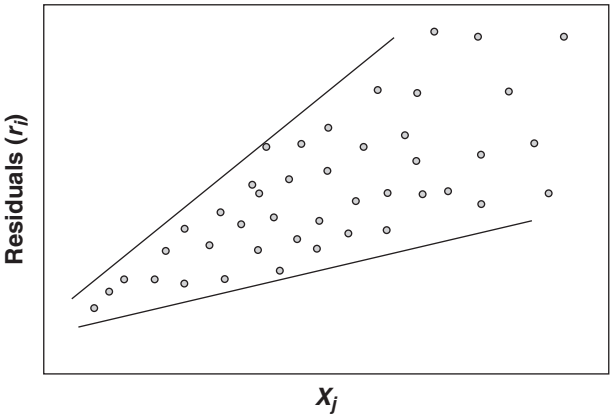



**Figure 4** Scatterplot of Residuals Versus *X* Exhibiting Heteroscedasticity

*t*-statistics are inflated, and regression parameters can turn out to be statistically significant, even though they are not. If the scatterplots of the residuals versus the explanatory variables show random scatters of points, no sign of heteroscedasticity is present (again, see Figure 2). When the scatterplots show that the band within which the residuals lie either increases or decreases with increasing *X* values, heteroscedasticity might be present. Figure 4 is an example of increasing residual variance.

### *Scatterplot of Studentized Residuals Versus Fitted Values*

Fitting the regression hyperplane to observed data establishes a linear relationship between the explanatory variables and the fitted values ($\hat{y}_i$), as indicated in Equation 2. Analogously, the standard regression assumptions can also be checked by plotting the standardized residuals against the fitted values ($\hat{y}_i$), rather than using the individual scatterplots of the standardized residuals versus the explanatory variables $(X_1, X_2, \ldots, X_p)$ as described above. However, the outcome of an analysis of the residual plots has to be the same by use of either the explanatory variables ($x_{ij}$) or the fitted values ($\hat{y}_i$). As a matter of fact, in the case of the simple regression analysis (i.e., only one explanatory variable *X*), the plots of the residuals versus *X* and versus $\hat{y}_i$ would be identical. Analogously, a pattern indicating heterogeneity would resemble a scatterplot as shown in Figure 4.

### *Scatterplot of Standardized Residuals Versus Index Numbers ("Index Plot")*

Residuals in regression analysis are supposed to be independent of one another. Especially in time series or spatial regression analysis, this assumption does not necessarily hold. Plotting the studentized residual against the index numbers, that is, plotting the residual in serial order, might be a first indication of whether the assumption of the independence of residuals holds. Again, a random pattern of the residuals indicates no violation of the regression assumptions, whereas a distinct pattern points toward the presence of auto- or spatial correlation. A case of severe autocorrelation is shown in Figure 5, where one clearly can identify a distinct pattern in the ordered residuals.

A second important application of the index plot of residuals is as a check for outliers. When using studentized residuals to identify outliers, reference is made to the outlyingness in *Y*-space, as residuals measure the vertical distances between the observations and the regression hyperplane. These outliers are not to be confused with outliers in *X*-space, which are usually identified using leverage values, that is, the entries on the main diagonal of the hat (projection) matrix. Though there is no standard rule on when one may call an

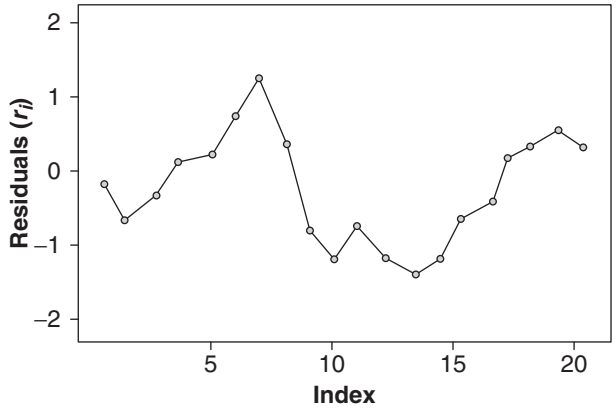



**Figure 5** Index Plot Exhibiting Correlated Residuals

observation an outlier, it is common practice to consider observations with studentized residuals that lie more than 2 to 3 standard deviations away from the mean of zero as outliers. For instance, Figure 6 shows two observations that can be considered outliers.

### *Potential Residual Plot*

The potential residual plot is a quick and convenient way of checking observations for their outlyingness in *X*-space (i.e., high-leverage points) and in *Y*-space (i.e., outliers) together in one scatterplot. The *potential* function defines an observation's outlyingness in *X*-space and the *residual* function

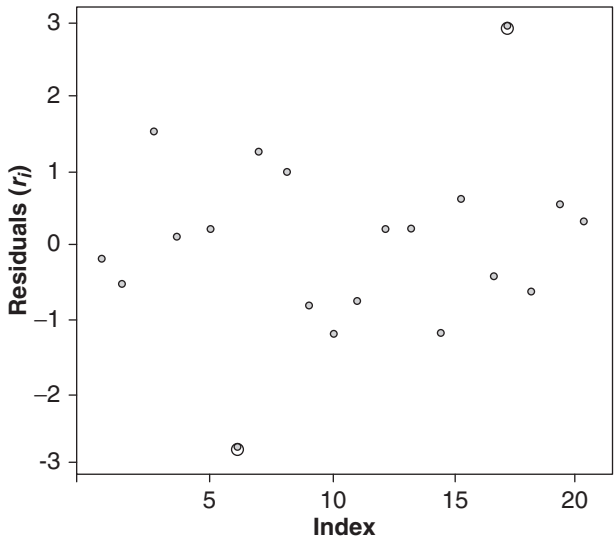



**Figure 6** Index Plot Exhibiting Two Potential Outliers

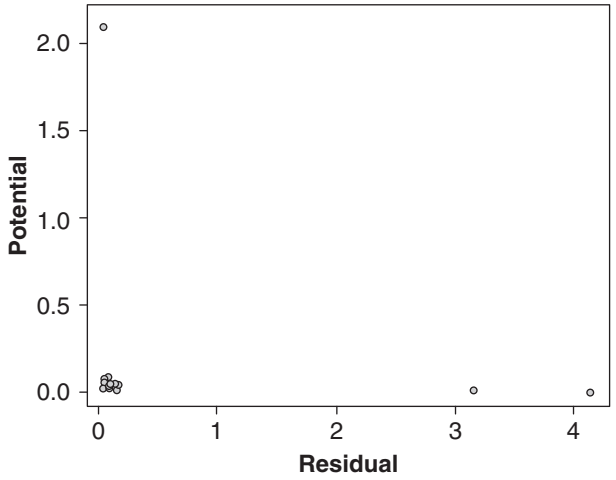



**Figure 7** Potential-Residual Plot With One High-Leverage Point and Two Outliers

analogously does the same for the *Y*-space. The potential and the residual functions are, respectively, defined as

$$\frac{p_{ii}}{1-p_{ii}} \quad \text{and} \quad \frac{p+1}{1-p_{ii}} \cdot \frac{d_i^2}{1-d_i^2}.$$

An example a potential residual function with one high-leverage point and two outliers is shown in Figure 7. For either form of outlyingness, further investigation of the corresponding observations is recommended to determine whether the observation is influential and whether further action is required. Influential observations are data points that uniquely influence the outcome of the estimated regression coefficients.

### *Partial Residual Plot*

The partial residual plot is well suited to situations in which the task is to graphically present the slope coefficient of $X_j$, namely $\hat{\beta}_j$. As shown in Figure 8, the slope of $\hat{\beta}_j$ is presented by plotting the residual + component against $X_j$, or

$$(e + \hat{\beta}_j X_j) \text{ versus } X_j.$$

The slope presented in the partial residual plot by all the individual points reflects the contribution of the explanatory variable $X_j$ on the fitted values $\hat{y}_i$. Note that each observation can be

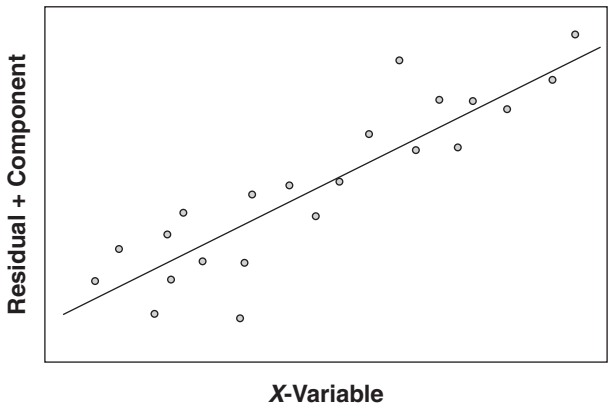



**Figure 8** Partial Residual Plot (Residual Plus Component Plot)

judged individually with respect to its own contribution toward the estimated regression coefficient $\hat{\beta}_j$. The straight line in Figure 8 is the equivalent to a fit of a simple regression model to the points in the partial residual plot.

The partial residual plot, or residual plus component plot, is useful in helping one visualize the marginal effect added variables have on the fitted values. Usually, the *t*-statistics in regression analysis are standard for deciding whether an explanatory variable should be retained or omitted from the model. In addition, the partial residual plot is useful in identification of whether the relationship between $Y$ and $X_j$ is linear. It is therefore suggested that the partial residual plot be used in conjunction with the *t* test.

### *Partial Regression Plot*

Also useful in determining the marginal effect on the regression model of adding an explanatory variable is the partial regression plot, or *added-variable plot*, a scatterplot of

$$Y\text{-residuals versus } X_j\text{-residuals.}$$

where the $Y$-residuals are obtained from regressing $Y$ on all explanatory variables except $X_j$, and the $X_j$-residuals are obtained by regressing $X_j$ on all other explanatory variables. The interpretation of the slope represented by the points is similar to that for the partial residual plot. In addition, partial regression plots are very useful in identifying high-leverage points and influential observations.

*Rainer vom Hofe*

***See also*** Residuals; Scatterplot

#### Further Readings

Chatterjee, S., & Hadi, A. S. (2006). *Regression analysis by example* (4th ed.). Hoboken, NJ: Wiley.
Greene, W. (2008). *Econometric analysis* (6th ed.). Upper Saddle River, NJ: Prentice Hall.
Gujarati, D. N. (2002). *Basic econometrics* (4th ed.). New York: McGraw-Hill.
Kennedy, P. (2008). *A guide to econometrics* (6th ed.). Hoboken, NJ: Wiley-Blackwell.
Tabachnik, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston: Allyn & Bacon.

## RESIDUALS

In statistics in general, the concept of residuals ($e$) is used when referring to the observable deviation of an individual observation in a sample to its corresponding sample mean. Not to be confused with the statistical error ($\varepsilon$), which refers to the often unobservable deviation of an individual observation from its often unknown population mean. The residual serves as an estimate for the unobservable statistical error when a sample is drawn from the population of interest but the population mean cannot be computed because the researcher simply does not have all individual observations belonging to the target population.

In this entry, the concept of residuals in regression analysis is defined, the use of residuals for checking the standard regression assumptions is discussed, and the role of the residuals in assessing the quality of the regression model and in hypothesis testing of the estimated regression coefficients is explained. In addition, the need for different types of residuals is described.

### Regression Analysis

The concept of residual in regression analysis is of utmost importance, particularly when one is

applying the regression method of ordinary least squares. In ordinary least squares regression analysis, the goal of which is to fit a line or (hyper)plane to observed data when one is using more than one explanatory variable, residuals depict the vertical distances between observed values and the corresponding fitted values. Conceptually, the residual is the part of the dependent variable ($Y$) that is not linearly related to the explanatory variables ($X_1, X_2, \ldots, X_p$). In other words, it is the part of $Y$ that cannot be explained by the estimated regression model.

In a regression model defined as

$$Y = X\beta + \varepsilon, \tag{1}$$

the $i$th fitted value, the value lying on the hyperplane, can be calculated as

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}; \quad i = 1, 2, \ldots, n, \tag{2}$$

where $p$ refers to the number of explanatory variables and $n$, the number of observations.

Analogously, the vertical distance between observed and fitted values for the $i$th observation, the $i$th residual ($e_i$), is defined as

$$e_i = y_i - \hat{y}_i, \tag{3}$$

which is graphically shown for the simple regression model in Figure 1.

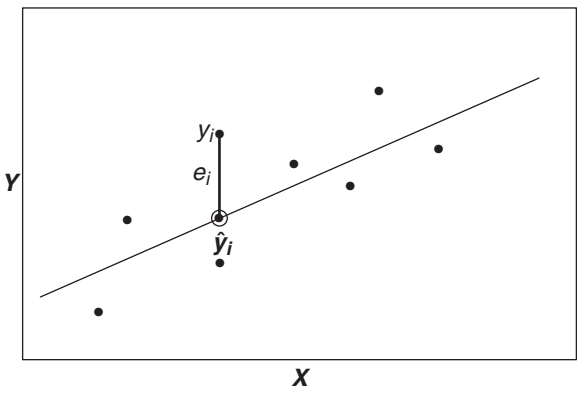



**Figure 1** Residual ($e_i$) in Ordinary Least Square Regression Analysis

## Standard Regression Assumptions

The assumptions about the error terms for the ordinary least square estimation are that the error terms are independently and identically distributed normal random variables, each with a mean of zero and a common variance, often expressed as

$$\varepsilon \sim NID(0, \sigma^2 I_n), \tag{4}$$

where $\sigma^2 I_n$ denotes a $(n \cdot n)$ matrix with the individual error variances on the main diagonal being equal to $\sigma^2$ and all the covariances on the off-diagonal being zero. For instance, for $n = 6$, the matrix $\sigma^2 I_n$ is

$$\begin{pmatrix} \sigma^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma^2 \end{pmatrix}.$$

Because of these error term assumptions, a standard practice in applied regression analysis is to include a series of tests and/or residual plots to check whether the residuals violate some of these assumptions. The literature is rich on statistical tests based on the calculated residuals, which include the Durbin–Watson test for autocorrelation (i.e., the errors are independent of each other) and the Breusch–Pagan test for heteroscedasticity (i.e., constant variance of the errors).

In addition to all these statistical tests, residual plots are an excellent and necessary tool to check for potential violations of the standard regression assumptions. As explained below, however, for better interpretation most residual plots do not use the original estimated residuals but rather use transformed residuals, such as standardized or studentized residuals. For instance, the normal probability plot of studentized residuals is suitable to check for the normality assumption, while the scatterplot of studentized residuals versus the explanatory variables is usable to check for the linearity and the constant variance assumptions. Further, the index plot of studentized residuals will point toward violations of the independence of errors, such as autocorrelation or spatial dependence of the error terms.

## Regression Model Quality and Hypothesis Testing

The importance of checking the assumptions about the error terms becomes more obvious when one considers that the residuals are not simply distance measures of how far individual data points are away from the regression line or in multiple regression analysis from the (hyper)-plane, but that they play an important role in assessing the quality of the estimated regression model as well as in hypothesis testing of the estimated regression coefficients. In either case, the standard error of the estimate $\hat{\sigma}^2$ needs to be calculated:

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-p-1}, \tag{5}$$

which in return is used in the calculation of the variances of the estimated regression coefficients:

$$Var(\hat{\beta}) = \hat{\sigma}^2 (X'X)^{-1}. \tag{6}$$

Taking the square roots of the variances of the estimated regression coefficients in Equation 6 in return gives us the corresponding standard errors of regression parameters:

$$se(\hat{\beta}_j) = \sqrt{Var(\hat{\beta}_j)} = \hat{\sigma}\sqrt{c_{jj}}, \tag{7a}$$

where $C = (X'X)^{-1}$. The standard errors can also be expressed as

$$se(\hat{\beta}_j) = \frac{\hat{\sigma}}{\sqrt{\sum (x_{ij} - \overline{x})^2}}; \quad i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, p. \tag{7b}$$

In regression analysis, the standard error is a common measure of the precision of the estimated parameters. Thus, the smaller the sum of squared residuals, the more precise are the estimated parameters.

In regression analysis, the most commonly used measure for the goodness of fit of the regression model is $R^2$, also referred to as the *coefficient of determination*. $R^2$ is defined as

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum e_i^2}{\sum (y_i - \overline{y})^2}, \tag{8}$$

where *SST* is the total sum of squared deviations in *Y*, *SSR* is the sum of squares explained by the regression model, and *SSE* is the sum of squared residuals. Although *SSR* is an indicator for the quality of chosen ($X_1$, $X_2$, . . . , $X_p$) as predictor variables, *SSE* is a measurement for the error in the prediction. This decomposition of the total sum of squared deviations in *Y* (*SST*) is part of the analysis of variance.

A second important statistic in regression analysis that relies on the sum of squared residuals (*SSE*) is the *F*-statistic:

$$F = \frac{SSR/p}{SSE/(n-p-1)} = \frac{\sum (\hat{y}_i - \overline{y})^2/p}{\sum e_i^2/(n-p-1)}. \tag{9}$$

Rather than examining the hypotheses of the individual regression coefficients, the $\beta$s, the *F*-statistic tests whether all the regression coefficients are simultaneously zero. An important feature of the *F*-statistic is its versatility when one is testing different scenarios. Assuming, for instance, that adding explanatory variables to the regression model will improve its overall predictive power, a significant decrease in the *SSE* would imply that the addition of explanatory variables improved the fit of the regression model. For this scenario, the *F*-statistic is calculated as

$$F = \frac{[SSE(RM) - SSE(FM)]/(p+1-k)}{SSE(FM)/(n-p-1)}, \tag{10}$$

where *SSE*(FM) = *SSE* of the full model, *SSE*(RM) = SSE of the reduced model, $p$ = number of explanatory variables in the full model, and $k$ = the number of explanatory variables in the reduced model.

As this example indicates, differences in the sums of squared residuals of the reduced versus the full model lead to various hypothesis tests. Besides the already described standard hypothesis test—whether all estimated coefficients are simultaneously zero—other commonly used *F*-tests check whether (a) a subset of the estimated parameters is zero, (b) some of the $\beta$s are equal to each other, and (c) some imposed constraints on

regression coefficients are being satisfied (e.g., $\beta_1 + \beta_2 = 1$).

## Types of Residuals

One important property of residuals in ordinary least square regression analysis is that they always sum to zero:

$$\sum e_i = 0. \tag{11}$$

In other words, the sum of the vertical distances above the fitted regression line, or the sum of all positive residuals, is equal to the sum of the vertical distances below the fitted regression line, or the sum of all negative residuals. The direct implication is that although the error terms are independent, the residuals are not. A second important property of the residuals is that they resemble only estimates of the true and unobservable error terms and as such do not have the same variance $\sigma^2$. For a direct comparison of residuals, such as for the detection of outliers, the residuals need to be standardized first. The variance of the $i$th residual is

$$Var(e_i) = \sigma^2(1 - p_{ii}), \tag{12}$$

where $p_{ii}$ is the $i$th leverage value. In multiple regression analysis, the $i$th leverage value is the $i$th diagonal element in the projection matrix, also known as *hat matrix*. The projection matrix is defined as

$$P = X(X'X)^{-1}X', \tag{13}$$

and its importance in regression analysis becomes more apparent when rewriting the vector of residuals as

$$e = Y - \hat{Y} = Y - PY = (I_n - P)Y. \tag{14}$$

Dividing the $i$th residual by its standard deviation then defines the standardized residual $z_i$:

$$z_i = \frac{e_i}{\sigma\sqrt{1 - p_{ii}}}, \tag{15}$$

which by definition has a mean of zero and a standard deviation of 1. The problem is that the standardized residual depends on the unknown standard deviation of the error terms $\sigma$. By use of the unbiased standard error of the estimate, the $i$th internally studentized residual $r_i$ is defined as

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - p_{ii}}}. \tag{16}$$

While the internally studentized residuals all have a standard deviation of 1, or they have the same variance, they no longer result in a sum of zero. Further, it should be noted that when one is usually referring to standardized residuals, the reference is often being made to internally studentized residuals.

Externally studentized residuals, on the other hand, use a slightly different unbiased estimate of the error term variance $\sigma^2$:

$$\hat{\sigma}^2_{(i)} = \frac{\sum e^2_{(i)}}{n - p - 2} = \frac{SSE_{(i)}}{n - p - 2}. \tag{17}$$

The difference between this and the commonly used estimate for $\sigma^2$ is that in this alternative estimate, the $i$th observation is excluded from the regression analysis when one is obtaining the $SSE_{(i)}$, as indicated by the subscript $(i)$. A second adjustment in the degrees of freedom in the denominator states that only $(n - 1)$ observations are used in the regression analysis. Externally studentized residuals are then defined as

$$r^*_i = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1 - p_{ii}}}. \tag{18}$$

The advantage of the externally studentized over the internally studentized residuals is that they follow a $t$-distribution with $n - p - 2$ degrees of freedom. However, for larger sample sizes, internally studentized residuals also approximate a standard normal distribution.

When the goal is to check the regression assumptions using residual plots, usually either version of residual is sufficient because one would use the residual plots only to identify a distinct pattern, which points toward one or the other model violation. In cases, however, in which a direct comparison between residuals is necessary, as for instance for the detection of outliers in $Y$-space, studentization of the residuals is necessary. It is a standard convention among researchers to consider observations with studentized residuals larger than 2 to 3 standard deviations away from the mean as potential outliers in the dependent

variable. To complete this discussion, outliers in the explanatory variables can be detected with leverage values, and observations with large leverage values are potential outliers in the explanatory variables. Note that leverage values by definition fall between 0 and 1, and observations with leverage values, that is, leverage values that are larger than twice the mean of all leverage values, should be marked for further investigation.

Once outliers in either the dependent or explanatory variables are identified, one can determine, using both the studentized residual, $r_i$, and the leverage value, $p_{ii}$, whether the $i$th observation is influential or not. In other words, would deleting the $i$th observation from the data set change significantly the estimated regression coefficients? The *Cook's distance measure*, one of the many influence measures available to check whether an outlier is influential, is defined as

$$C_i = \frac{r_i^2}{p+1} \cdot \frac{p_{ii}}{1-p_{ii}}; i = 1, 2, \ldots, n, \qquad (19)$$

where the first term of the product defines the outlyingness of the $i$th observation in $Y$-space based on its internally studentized residual, and the second term, also referred to as the *potential function,* defines the outlyingness of the $i$th observation in $X$-space based on the corresponding leverage value. An index plot of Cook's distance measures is a first step in identifying large $C_i$ values that stand visually apart from the other distance measures and that indicate that deleting the $i$th observation from the data set will significantly change the outcomes of the regression results.

### Final Thoughts

Although regression analysis focuses on obtaining robust, unbiased, and meaningful regression coefficients that show the functional linear relationship between the response variable and a set of explanatory variables, an equal amount of attention should be paid to the residuals. In fact, residuals are the main source of information when one is checking the regression assumptions, conducting hypothesis tests, or assessing the quality of the estimated regression model.

*Rainer vom Hofe*

***See also*** Regression Artifacts; Regression Coefficient; Regression Discontinuity

#### Further Readings

Chatterjee, S., & Hadi, A. S. (2006). *Regression analysis by example* (4th ed.). Hoboken, NJ: Wiley.
Greene, W. (2008). *Econometric analysis* (6th ed.). Upper Saddle River, NJ: Prentice Hall.
Gujarati, D. N. (2002). *Basic econometrics* (4th ed.). New York: McGraw-Hill.
Kennedy, P. (2008). *A guide to econometrics* (6th ed.). Hoboken, NJ: Wiley-Blackwell.
Tabachnik, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston: Allyn & Bacon.

## Response Bias

The response to a survey question is said to be biased if it does not reveal, on average, the true value of the target variable. This definition is closely related to the definition of a bias in statistics: The bias of an estimator is the difference between its expected value (when computed for a given sample) and the true value of the corresponding parameter (in the population). Biased responses may arise for many types of target variables: attitudes, behaviors, preferences, and expectations; sociodemographic characteristics such as age, level of education, or labor-market status; frequencies such as the number of purchases of a good or the number of doctor visits in a specified period; and monetary quantities such as income, financial assets, or consumption expenditure. This entry examines sources and implications of response bias and presents strategies for dealing with biased responses.

### Sources of Response Bias

For several decades, psychologists, sociologists, and survey researchers have worked to understand the cognitive and communicative processes that generate survey responses. A central insight is that answering a survey question is a complicated process consisting of several distinct tasks (even though these tasks are not completely sequential or independent processes). Biased responses may

arise at any of these stages. It is important to note that there is no single process that would be responsible for response bias. Potential sources of biased responses can be illustrated with a conceptual model of the survey response process that distinguishes four stages.

- First, the respondent needs to comprehend the question: He or she needs to understand what information is sought. Poor wording may easily lead to systematic misunderstanding of the question and thus to biased responses.
- Second, the respondent must retrieve the relevant information from his or her memory. This stage activates retrieval strategies that often require filling in missing details. The success of these retrieval strategies depends on many variables, such as the respondent's cognitive ability and memory capacity.
- The third stage involves making a judgment: Is the information just retrieved from memory complete? If not, the respondent uses estimation strategies that integrate the retrieved information with other salient information that is contained in the survey questionnaire, provided by the interviewer, and so on. These estimation strategies often involve a notion of *satisficing*, that is, the respondent uses just as much time and effort as is needed to construct a response that is "good enough." Estimation strategies and the corresponding heuristics typically result in biased responses rather than random errors.
- The final stage is to report the response that has been constructed in the first three stages. The respondent may alter his or her response because he or she does not want to reveal the constructed response for reasons of confidentiality or social desirability.

## Examples

### *Measurement of Attitudes and Behaviors*

Many survey measurements refer to attitudes or behaviors that may or may not be socially desirable. For instance, in a country in which the death penalty is considered immoral, respondents who are in favor of it may nevertheless report that they oppose it. Other examples are questions on substance use or on donations to charities. Depending on the circumstances, social desirability will result, on average, in under- or overreporting of attitudes or behaviors. Such effects are likely to be stronger in personal interviews than in self-administered mail or Internet surveys.

### *Frequency Reporting*

When respondents are asked to report the frequency of events, they use a variety of estimation strategies: They count salient episodes; they base their estimate on an interval-specific rate; or they rely on cues about typical frequencies that are provided by the response categories. Depending on the context (e.g., the length of the period for which a frequency is to be reported), these strategies may result in systematic over- or underreporting.

### *Expenditure Measurement*

Household surveys often ask respondents to report the amount spent on a certain class of consumption goods in a previous period. Since such quantities are not directly represented in a respondent's memory, the respondent must use an estimation strategy whose accuracy depends on many variables: how well the target quantity is described in the question, how easily typical purchases can be retrieved from memory, and so on. Respondents may integrate salient features of the questionnaire, such as numerical clues provided by the response categories. The errors introduced by such estimation strategies are nonrandom and lead to biased responses.

## Statistical Implications of Response Bias

When responses to a survey question are biased, the measured variable is subject to measurement error (this is only one of several forms of survey error, the others being coverage error, sampling error, and nonresponse error). Crucially, measurement error that stems from biased responses does not conform to the paradigm of classical measurement error, which assumes that errors are random, that is, statistically independent of the true value of the measured variable (and of other variables in a statistical model). In fact, it is hard to imagine a realistic response process that would generate a survey response that meets the classical measurement error assumptions.

The statistical analysis of data that are subject to response bias requires the specification of an explicit, nonclassical error model that, in turn, requires making assumptions about the process that generates the bias. Typically, such assumptions are untestable with the available data.

### Strategies for Dealing With Biased Responses

The best strategy to deal with biased responses is to avoid them in the first place. Consulting the established literature on survey response behavior helps one avoid well-known potential pitfalls. Often, the existing literature will suggest that a specific question or other survey design feature may introduce response bias, but it may be difficult to judge *ex ante* whether in a specific situation a potential bias will in fact be large enough to affect the subsequent analysis of the data and the substantive conclusions one may draw. In these situations, a researcher may vary the wording, framing, or other design features experimentally—either in a pretest or in the main study. Such survey experiments allow the researcher to test whether a bias actually exists, that is, whether response distributions differ between treatments. If a significant bias exists, the researcher can either adjust the design to avoid the bias or, in case this is not possible, deal with the bias at the analysis stage, using advanced statistical methods and explicit error models.

*Joachim K. Winter*

***See also*** Differential Item Functioning; Internal Validity; Sampling Error; Survey; Systematic Error; Validity of Measurement

### Further Readings

Krosnick, J. A., & Fabrigar, L. R. (in press). *The handbook of questionnaire design.* New York: Oxford University Press.

McFadden, D., Bemmaor, A., Caro, F., Dominitz, J., Jun, B., Lewbel, A., et al. (2005). Statistical analysis of choice experiments and surveys. *Marketing Letters, 16,* 183–196.

Schaeffer, N. C., & Presser, S. (2003). The science of asking questions. *Annual Review of Sociology, 29,* 65–88.

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers.* San Francisco: Jossey-Bass.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response.* New York: Cambridge University Press.

## Response Surface Design

A response curve is a mathematical function representing the relationship between the mean of a response variable and the level of an explanatory variable. The curve might be linear, a higher degree polynomial, or some nonlinear function. A response surface is a mathematical function representing the relationship between the mean of a response variable and the levels of several explanatory variables. The surface might be linear, a higher order polynomial, or a nonlinear function. In an experiment, the explanatory variables, or treatment factors, are under the control of the experimenter, and the level of each must be chosen for each experimental run. A response surface design is the set of combinations of the levels of the explanatory variables used in a particular experiment.

There are many reasons experimenters might want to fit response surface models, such as (a) to gain a simple mathematical and graphical description of the effects of the treatment factors on the response, (b) to identify the factors that have important effects on the response, (c) to estimate the effects of changing the levels of the factors from standard operating conditions, or (d) to enable estimation of the levels of the factors that optimize (maximize, minimize, or achieve a target level of) the response. The use of economical response surface designs can help researchers answer research questions precisely with considerably less time and cost than more ad hoc methods of experimentation. Response surface designs are particularly useful when combined with the experimenters' expert knowledge and previous results.

In many areas of research, especially applied research, questions arise that can be answered by experimenting to determine the effects of several factors on one or more responses of interest. In such cases, the use of factorial-type experiments, in which all factors are studied in the same experiment, have long been known to be the most economical way of experimenting. In standard

factorial designs, it is usually assumed that the factors are qualitative, or at least that their effects will be modeled in terms of main effects and interactions, ignoring the quantitative nature of the levels.

In contrast, if the factors represent continuous variables, it is natural to fit a response surface model. Occasionally, mechanistic information will be available to determine the specific response surface function that should be used, but more often a purely empirical polynomial response surface model will be used to approximate the unknown function. Statistical methods for empirical modeling, collectively known as *response surface methodology,* were developed by George Box and his colleagues at Imperial Chemical Industries, in England, in the 1950s and have been further refined since then. This entry describes useful designs for fitting and checking first-order and second-order models and then considers complications that can arise.

## First-Order Designs

### *Factorial Designs*

Although there are several other designs for fitting first-order, that is, multiple linear regression, models, the only ones used in practice are two-level factorial designs, or *fractional factorial designs.* The two levels of a factor, usually coded $-1$ and $+1$, represent low and high levels of that factor. For example, with three factors, the following design, presented in a standard order, might be used.

| $X_1$ | $X_2$ | $X_3$ |
|---|---|---|
| −1 | −1 | −1 |
| −1 | −1 | +1 |
| −1 | +1 | −1 |
| −1 | +1 | +1 |
| +1 | −1 | −1 |
| +1 | −1 | +1 |
| +1 | +1 | −1 |
| +1 | +1 | +1 |

The usual advice is to use levels that are far enough apart that experimenters can be confident of identifying any practically meaningful effects, but close enough that a low-order polynomial model can be expected to give a reasonable approximation to the response surface. The order of the runs should be randomized before the experiment is run, and the levels of each factor reset for each run. The model can be written as

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q + \varepsilon,$$

where $Y$ represents the response, $x_1, \ldots, x_q$ represent the coded levels of the factors, and $\varepsilon$ is a random error term with mean zero, representing deviations of individual observations from the mean. The error terms are usually assumed to have constant variance and to be independent and are often assumed to have normal distributions. Data analysis proceeds exactly as for any multiple regression model except that the coded levels of variables are usually used.

Factorial designs are particularly useful for identifying a small number of important factors out of many potentially important ones and for identifying new levels of the factors that should be explored in future experiments. For example, if the high level of one factor gives a much better response than the low level, it might be fruitful to explore some even higher levels.

### *Pure Error and Lack of Fit*

Although two-level factorial designs allow the first-order model to be fitted efficiently, they do not allow the appropriateness of this model to be assessed. A simple modification, which allows this, is to include a few center points (i.e., runs in which each factor is set at a level halfway between the low and high levels). Often, only 3 to 5 center points are run. The design, again in a standard order, then becomes the following:

| $X_1$ | $X_2$ | $X_3$ |
|---|---|---|
| −1 | −1 | −1 |
| −1 | −1 | +1 |
| −1 | +1 | −1 |
| −1 | +1 | +1 |
| +1 | −1 | −1 |
| +1 | −1 | +1 |
| +1 | +1 | −1 |
| +1 | +1 | +1 |
| 0 | 0 | 0 |
| 0 | 0 | 0 |
| 0 | 0 | 0 |
| 0 | 0 | 0 |

The run order should again be randomized, and the levels of each factor reset for each run. In particular, the center points must be genuine replicates, fully reset for each run, and not just repeated measurements from the same run. The variance of the responses from the center points then measures the run-to-run variance in the absence of treatment differences, often called *pure error*. This design allows more reliable hypothesis tests of the terms in the response surface model. It also allows the difference in response between the factorial points and the center points, which measure curvature in the response surface, to be used to assess the lack of fit of the first-order model. A hypothesis test that shows lack of fit indicates that the first-order model is inadequate and that one should consider doing more experimentation to fit a model of second order.

## Second-Order Designs

### *Three-Level Designs*

In order to fit the second-order polynomial model,

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q + \beta_{11} x_1^2 + \cdots + \beta_{qq} x_q^2 + \beta_{12} x_1 x_2 + \cdots + \beta_{(q-1)q} x_{q-1} x_q + \varepsilon,$$

at least three levels of each factor are used, and it is most efficient to use equally spaced levels. The most commonly used second-order response surface design is the central composite design, which includes factorial points and center points, as above, and other so-called axial points, in which one factor is at its high or low level and all other factors are at their middle levels. For three factors, the central composite design is as follows:

| $X_1$ | $X_2$ | $X_3$ |
|---|---|---|
| −1 | −1 | −1 |
| −1 | −1 | +1 |
| −1 | +1 | −1 |
| −1 | +1 | +1 |
| +1 | −1 | −1 |
| +1 | −1 | +1 |
| +1 | +1 | −1 |
| +1 | +1 | +1 |
| −1 | 0 | 0 |
| +1 | 0 | 0 |
| 0 | −1 | 0 |
| 0 | +1 | 0 |
| 0 | 0 | −1 |
| 0 | 0 | +1 |
| 0 | 0 | 0 |
| 0 | 0 | 0 |
| 0 | 0 | 0 |
| 0 | 0 | 0 |
| 0 | 0 | 0 |
| 0 | 0 | 0 |

The model is easily fitted by least squares by a standard multiple regression program. The central composite design is very popular because it uses relatively few runs and is reasonably efficient and because it allows sequential experimentation since it can be obtained by the addition of the axial points, and usually some additional center points, to the two-level factorial design.

A useful class of three-level designs when experimentation is very expensive or time-consuming is the small composite designs, which replace the factorial portion of the central composite design with a small fractional factorial.

| $X_1$ | $X_2$ | $X_3$ |
|---|---|---|
| −1 | −1 | −1 |
| −1 | +1 | +1 |
| +1 | −1 | +1 |
| +1 | +1 | −1 |
| −1 | 0 | 0 |
| +1 | 0 | 0 |
| 0 | −1 | 0 |
| 0 | +1 | 0 |
| 0 | 0 | −1 |
| 0 | 0 | +1 |
| 0 | 0 | 0 |

Another popular class of designs are the Box–Behnken designs, which have very specific structures and are available in several books and packages. With three factors, the Box–Behnken design is as follows:

| $X_1$ | $X_2$ | $X_3$ |
|---|---|---|
| −1 | −1 | 0 |
| −1 | +1 | 0 |
| +1 | −1 | 0 |
| +1 | +1 | 0 |
| −1 | 0 | −1 |
| −1 | 0 | +1 |
| +1 | 0 | −1 |
| +1 | 0 | +1 |
| 0 | −1 | −1 |
| 0 | −1 | +1 |
| 0 | +1 | −1 |
| 0 | +1 | +1 |
| 0 | 0 | 0 |
| 0 | 0 | 0 |
| 0 | 0 | 0 |

One advantage of the Box–Behnken design over the central composite design is that it does not use points with extreme levels of all factors simultaneously. This can be advantageous when such points might be expected to give very untypical responses or be dangerous or unethical to use.

All the designs described so far emphasize the estimation of the second-order model with relatively few experimental runs. This is appropriate when experimentation is costly and run-to-run variation is small, such as with many physical systems. In other types of experiment, especially when biological materials are being used, run-to-run variation will be considerably higher. It might then be advantageous to use somewhat larger designs in order to improve the estimation of the response surface. In such cases the class of subset designs can be useful. These are made up of the subsets $S_0, \ldots, S_q$, where $S_r$ is the set of points with $r$ factors at $\pm 1$, the other $q - r$ factors being at 0. By combining different numbers of replicates of these subsets, researchers can find many useful designs. With three factors, the subsets are as follows:

| $S_3$ | | | $S_2$ | | | $S_1$ | | | $S_0$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $X_1$ | $X_2$ | $X_3$ | $X_1$ | $X_2$ | $X_3$ | $X_1$ | $X_2$ | $X_3$ | $X_1$ | $X_2$ | $X_3$ |
| −1 | −1 | −1 | −1 | −1 | 0 | −1 | 0 | 0 | 0 | 0 | 0 |
| −1 | −1 | +1 | −1 | +1 | 0 | +1 | 0 | 0 | | | |
| −1 | +1 | −1 | +1 | −1 | 0 | 0 | −1 | 0 | | | |
| −1 | +1 | +1 | +1 | +1 | 0 | 0 | +1 | 0 | | | |
| +1 | −1 | −1 | −1 | 0 | −1 | 0 | 0 | −1 | | | |
| +1 | −1 | +1 | −1 | 0 | +1 | 0 | 0 | +1 | | | |
| +1 | +1 | −1 | +1 | 0 | −1 | | | | | | |
| +1 | +1 | +1 | +1 | 0 | +1 | | | | | | |
| | | | 0 | −1 | −1 | | | | | | |
| | | | 0 | −1 | +1 | | | | | | |
| | | | 0 | +1 | −1 | | | | | | |
| | | | 0 | +1 | +1 | | | | | | |

The best designs depend on the number of runs that can be done and the precise objectives of the experiment. For example, with 30 runs, the design $2S_3 + S_2 + 2S_0$ gives good estimation of the linear and interaction coefficients and of pure error, but not such good estimation of the quadratic coefficients, whereas $S_3 + S_2 + S_1 + 4S_0$ gives better estimation of the quadratic coefficients but allows only 3 degrees of freedom for pure error.

### *Designs With More Levels*

Although three levels are enough to fit the second-order model and estimate pure error, they do not allow a very good test for lack of fit, since at least four levels are needed to estimate pure cubic effects. Four-level designs are sometimes used, but much more commonly the central composite design is adapted to have five levels by placing the axial points farther out than $\pm 1$. Any value can be used for the axial points, but a popular choice is $\sqrt{q}$ so that, in terms of the coded factor levels, the region of experimentation becomes a sphere, or hypersphere. With three factors, for example, a central composite design is

| $X_1$ | $X_2$ | $X_3$ |
|---|---|---|
| −1 | −1 | −1 |
| −1 | −1 | +1 |
| −1 | +1 | −1 |
| −1 | +1 | +1 |
| +1 | −1 | −1 |
| +1 | −1 | +1 |
| +1 | +1 | −1 |
| +1 | +1 | +1 |
| $-\sqrt{3}$ | 0 | 0 |
| $+\sqrt{3}$ | 0 | 0 |
| 0 | $-\sqrt{3}$ | 0 |

| | | |
|---|---|---|
| 0 | $+\sqrt{3}$ | 0 |
| 0 | 0 | $-\sqrt{3}$ |
| 0 | 0 | $+\sqrt{3}$ |
| 0 | 0 | 0 |
| 0 | 0 | 0 |
| 0 | 0 | 0 |
| 0 | 0 | 0 |
| 0 | 0 | 0 |
| 0 | 0 | 0 |

Small composite designs and subset designs can be modified in a similar way, but with subset designs it is always necessary to include at least one center point.

### *Sequential Design*

It was noted above that the central composite design lends itself to sequential experimentation, with the factorial points and some center points being run first to fit and check the first-order model. If this seems inadequate, then the axial points and some more center points are added to allow the second-order model to be fitted and checked. If, however, the first-order model seems adequate, the first experiment might suggest that different levels of the factors should be used; for example, higher responses can be obtained outside the region of experimentation. In such cases, the next experiment should probably be run in a new region of experimentation. For example, if, in the original first-order design, it seemed that the response at $(+1, +1, -1)$ was much higher than anywhere else in this experimental region, a new first-order design with levels $(+1, +3)$ for $X_1$ and $X_2$ and $(-3, -1)$ for $X_3$ could be run. There are many other possibilities; for example, the levels of some factors could remain the same, the range of levels could be expanded or contracted, or some factors could be dropped from the experiment.

If the results of the initial experiment suggest that the optimal response might be achieved far from the region of experimentation, then some experiments could be run along the direction of steepest ascent. For example, if the estimated coefficients for the three factors' linear effects are $-2$, 1, and 3, respectively, runs could be made at $(-1, 0.5, 1.5)$, $(-2, 1, 3)$, $(-3, 1.5, 4.5)$, and so on. These can either be made as a batch, possibly with some replication, or one at a time. This technique was popular in the early days of response surface methodology but seems to be used less now. It probably works best if run-to-run variation is very small.

### *Optimal Design Criteria*

An approach to choosing a response surface design that is different from the standard designs described above is to search for a design that optimizes some design criterion. An obvious criterion is to minimize the average variance of the parameter estimates, and this is known as *A-optimality.* A computationally simpler, and hence more popular, criterion is *D-optimality,* which minimizes the volume of a joint confidence region for the parameters. Other criteria are based on minimizing the average (*I-optimality*) or maximum (*G-optimality*) variance of the estimated response in the region of experimentation.

Choosing an optimal design usually involves specifying a large set of candidate points, such as the full three-level factorial design, and running a search algorithm to find the combination of these that optimizes the relevant criterion. Most of these combinations work by starting with a randomly chosen design and then systematically exchanging points in the current design with points in the candidate set, accepting any exchange that improves the design. Such algorithms are available in many statistical software packages.

There is no agreement among statisticians or users about whether it is better to use an optimal design than one of the standard designs. In theory, it would be best to define an optimality criterion that exactly encapsulates all the objectives and limitations of the experiment and then choose a design that will optimize this criterion. However, it is probably impossible to define such an exact criterion in practice, or to do so would cost more time and effort than doing the experiment. This is why many experimenters prefer to use standard designs that are known to have many good properties.

## Blocking

Up to now this discussion has assumed that the run order would be completely randomized within each experiment and that run-to-run variation would be used to assess the effects of factors.

Sometimes, however, it is known or expected before the experiment that there will be systematic variation between runs, irrespective of which treatments are applied. For example, runs made on the same day might tend to be more similar than runs made on different days. Then the technique of blocking can be used in order to separate variation between days from variation between runs on the same day. The advantage of this separation is that the effects of the factors can then be assessed relative to the variation between runs made on the same day so that more precise estimates of these effects can be made.

Unfortunately, unlike simple treatments or factorial designs, theory provides no simple mathematical method for arranging response surface designs in blocks. Instead, one usually has to use a search algorithm with some optimality criterion. Again different criteria can be used, although the choice of criterion is less crucial for arranging a design in blocks than it is for searching for optimal treatment designs. Typical algorithms start with a random allocation of treatments to blocks and then systematically interchange them, accepting any interchanges that improve the criterion. Several algorithms have been suggested in recent years, and these have now started to appear in statistical software packages.

Sometimes there might be two (or more) types of blocks that can be used, either such that they are nested, such as days within weeks, or such that they are crossed, such as days and machines. Interchange algorithms can easily be extended to such nested block designs and row–column designs.

### Multistratum Designs

It is very common in industrial and engineering experiments to have some factors whose levels are harder to set than others. For example, it might be convenient to set the temperature just once every day, whereas other factors can be varied in several runs within each day. This randomization restriction is then very similar to that used in the split-plot factorial design, but again the designs do not have such a simple mathematical structure. The same idea can be extended to having factors at more than two levels of variation. The general class of such designs is called *multistratum response surface designs*. This is an area of much current research. Several algorithms and strategies for constructing these designs have been suggested but are not yet readily available in widely used software packages.

*Steven G. Gilmour*

***See also*** Block Design; Factorial Design; Multiple Regression; Polynomials; Random Assignment; Split-Plot Factorial Design

#### Further Readings

Atkinson, A. C., Donev, A. N., & Tobias, R. (2007). *Optimum experimental designs, with SAS*. Oxford, UK: Oxford University Press.

Box, G. E. P., & Draper, N. R. (2007). *Response surfaces, mixtures and ridge analyses* (2nd ed.). New York: Wiley.

Box, G. E. P., Hunter, J. S., & Hunter, W. G. (2006). *Statistics for experimenters* (2nd ed.). New York: Wiley.

Box, G. E. P., & Wilson, K. B. (1951). On the experimental attainment of optimum conditions (with discussion). *Journal of the Royal Statistical Society, Series B, 13,* 1–45.

Gilmour, S. G. (2006). Response surface designs for experiments in bioprocessing. *Biometrics, 62,* 323–331.

Goos, P. (2002). *The optimal design of blocked and split-plot experiments*. New York: Springer.

Khuri, A. I. (Ed). (2006). *Response surface methodology and related topics*. New York: World Scientific.

Khuri, A. I., & Cornell, J. A. (1996). *Response surfaces* (2nd ed.). New York: Dekker.

Myers, R. H., & Montgomery, D. C. (2002). *Response surface methodology* (2nd ed.). New York: Wiley.

Wu, C. F. J., & Hamada, M. (2000). *Experiments*. New York: Wiley.

## Restriction of Range

*Restriction of range* is the term applied to the case in which observed sample data are not available across the entire range of interest. The most common case is that of a bivariate correlation between two normally distributed variables, one of which has a range less than that commonly observed in the population as a whole. In such cases the observed correlation in the range restricted sample

will be attenuated (lower) than it would be if data from the entire possible range were analyzed.

The context in which restriction of range is discussed most often is that of criterion-related validity evidence of predictors in a selection context such as employee selection or educational institution admissions. For instance, common cases include those in which job applicants take a test of some type, and then only a subset of those applicants is hired. When the selection decision is based on the test scores, the range of the sample will be restricted.

## How Range Restriction Affects Correlation Coefficients

Figure 1 illustrates the effect of restriction of range on observed correlations. In Figure 1(a), data for 50 persons appears for both a predictor and a criterion. With this full sample, the correlation is .82. Figure 1(b) illustrates the effect of restriction of range on the predictor. Such a scenario is a common one in validation research that uses predictive validity designs in which there is a large pool of job or college applicants, yet only a subset of applicants with the highest scores on the predictor is selected. In this example, only persons with the top 15 scores on the predictor were selected. Thus criterion data would be available only for these persons. The observed correlation is attenuated from .82 in the full sample to .58 in the range-restricted sample.

Conversely, Figure 1(c) illustrates the effect of restriction of range on the criterion in which only the 15 persons with the highest criterion scores are included. Such a scenario is often encountered in validation research in which a concurrent validation design is employed. In such a design, current employees are given a predictor measure of interest and the correlation between predictor scores and the criterion is assessed. In this case, poorly performing employees likely would have been fired or would have quit their jobs, thus restricting the range of the performance criterion. In this example, the observed correlation is .51.

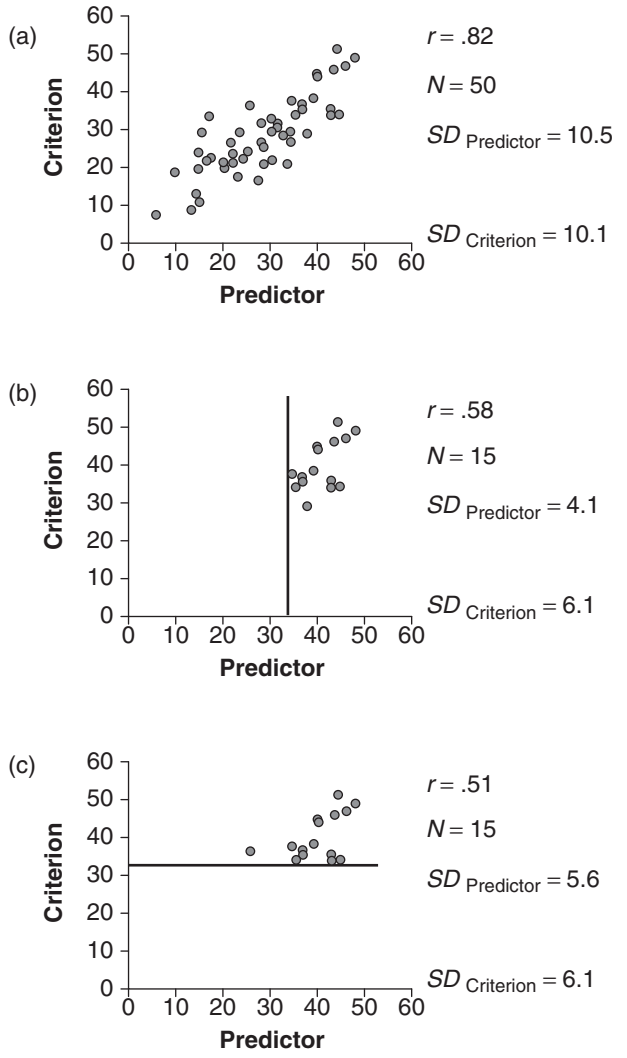



**Figure 1** Effects of Restriction of Range

## Types of Restriction of Range

Most researchers recognize a distinction between different types of restriction of range. The most obvious type is known as *direct* (or explicit selection) restriction of range. Direct restriction of range is that depicted in Figure 1(b), in which there is a predictor of interest that is given to a set of applicants and only those with the highest scores on the predictor are selected. In this case, the restriction of range is direct and purposeful. Direct restriction of range is witnessed in predictive validation designs in which the predictor data are collected, persons are selected into the organization based on the predictor, and later a criterion performance measure is collected. Because of the effect of restriction of range on the correlation between the predictor and the criterion, the American Educational Research Association, American Psychological

Association, and the National Council on Measurement in Education recommend that the predictor of interest not be used for selection until the validation process is complete.

A second type of restriction of range is *indirect* (or incidental selection) restriction of range. Indirect restriction of range occurs in either predictive or concurrent validation designs. In these cases, the predictor of interest is not used to make selection decisions in the organization; however, the predictor of interest is correlated with another predictor that is used for selection decisions. In the example of a predictive validation design, both the predictor of interest and a correlated other predictor are given to a set of applicants. Only those applicants with the highest score on the other predictor are selected; however, because of the correlation between the predictors, persons selected will tend to be those with the highest scores on the predictor of interest. In the concurrent validation design, a predictor of interest is given to existing employees. If those current employees were selected on the basis of their scores on another predictor that correlates with the predictor being validated, only employees with scores in the upper range of the predictor of interest will be available.

Note that formulas are available to estimate the unrestricted correlation coefficient given a range-restricted sample:

$$r_c = \frac{r(\sigma_u/\sigma_r)}{\sqrt{1 - r^2 + r^2(\sigma_u^2/\sigma_r^2)}},$$

where $r_c$ is the corrected correlation, $r$ is the observed correlation in the range-restricted sample, and $\sigma_u$ and $\sigma_r$ are the standard deviation of the predictor in the unrestricted and range-restricted samples, respectively.

*Adam W. Meade*

***See also*** Correlation; Range; Raw Scores

**Further Readings**

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

## Results Section

The purpose of the results section of a research paper is to present the key results of a research study without interpreting their meaning. The results should be presented (a) in an orderly sequence so that their sequencing tells a story and (b) where appropriate, with the use of summary techniques and visual aids, such as tables and figures. The results section does not need to include every result obtained or observed in a study. It presents only those results that are relevant to the objectives and the research questions of the study.

Usually the results section is not combined with the discussion section unless specified by the journal to which the research paper will be submitted. The results section in general should not refer to the discussion section, which has not yet been written (but the discussion section should frequently refer to the results section). The results section, however, should closely reflect the methods section. For every result there must be a method in the methods section.

Below is a checklist of the items to be included in a results section:

*Text* First, the results must be organized in an orderly sequence, such as chronological order, logical order, or in order of most to least important. One recommended order in the results section is as follows: (a) Summary of what the data set "looks like," with descriptive statistics (e.g., mean and standard deviation) for the relevant variables as found in preliminary analyses. (b) Examination of the hypotheses through further analyses. (c) Verification of which findings were or were not in the predicted direction, based on confirmatory analyses.

Second, whether the results are best presented in the form of text, tables, figures, or graphs is determined. Tables are good for including large amounts of information, especially of a repeated or comparative nature, for completeness. Scatterplots or histograms can provide a real feel for the data (including the noise). With complex data sets, bar graphs are useful so that the reader is not

overwhelmed with too many data points at once. When results are presented in tables or figures, the text should describe only the highlights of the findings and point the reader to the relevant tables and figures. The text should provide background information for the tables and figures, as well as observations not presented in the tables and figures. In other words, the text should complement the tables and figures, not repeat the same information.

Third, the text—not the statistics—should describe the patterns in the data. Statistics should not be used as though they were parts of speech. For example, "the correlation between variables A and B was $r = -.26$, $p = .01$" should be translated into words, with the statistics provided as evidence for the reported findings; for example, "a negative correlation was found between variables A and B, indicating that an increase in A was associated with a decrease in B ($r = -.26$, $p = .01$)." The word *prov*e should be used with caution. Because statistical tests are based on probability and can be in error, they do not really prove anything. The meaning of the alpha level or the null hypothesis need not be elaborated if the intended audience of a research paper is the scientific community; in this case, the researcher can assume the reader will have a working knowledge of statistics. The word *significant* should also be used with care. A statistically significant difference is one in which the probability that the groups of data are identical with respect to the characteristic under study is so low (e.g., less than .05) that the investigator can confidently conclude that the difference exists. The opposite is "statistically nonsignificant." A clinically significant difference is one that is considered important. The opposite is "clinically insignificant." The sentence "It is significant that we found no significant differences among the groups studied" is a valid, though perhaps confusing, statement.

*Tables and Figures* When used correctly, these are good summary techniques and visual aids to help describe the results. The common mistakes, however, are that tables and figures are too complex and too many. The following are several techniques to improve the presentation of tables and figures. First, tables and figures must be linked to the text in the results section. A citation in the text, such as "See Table 1" or "Figure 2 shows that . . ." should always be included in the results section for each table and figure. Second, each table and figure should "stand alone," that is, a reader should be able to understand each one without having to refer to the text. For example, each table and figure should have a comprehensive heading that describes its content (including the country, name, and year of the study if applicable); the number of cases or sample size to which the table or figure refers should be given; and any abbreviations used in the figure or table should be defined in a footnote, even if they are explained in the text. Third, tables and figures must be linked to the methods section. All statistical techniques used to generate tables and figures must be described adequately in the methods section. Finally, approval for reproduction of any material (including figures and tables) published elsewhere should be obtained from the copyright holders and/or authors before submission of the manuscript to a journal. Authors usually are responsible for any copyright fee involved.

The following are some general technical guidelines.

*For tables and figures:*

- Check the journal guidelines on the maximum number of tables or figures allowed, and find out whether the journal uses color (often there is a charge) or black and white.
- Number tables and figures consecutively in the same sequence they are first mentioned in the text.
- Number tables and figures in separate sequences; for example, Table 1, 2, and so on; Figure 1, 2, and so on.
- Do not put a box around tables and figures.
- Avoid decorative background, such as shading or patterned bars.
- Depending on the journal, tables and figures should be placed at the end of the report, after the references section (some journals require authors to indicate in the text of the results section the approximate location of the tables and figures), or located appropriately within the text of the results section.

*For tables:*

- Many journals allow only horizontal lines and no vertical line in tables.
- Use horizontal lines to separate rows only if the rows deal with different types of variables, such as age, income, and education.

- Make sure that all numbers, especially if they have decimal points, line up properly.
- Make sure that the percentages of the categories of a variable add up to 100; round all numbers, except *p* values, to two decimal places.
- To indicate that some numbers represent statistically significant differences, give the test used and *p* values or confidence intervals.
- Give column and row totals where appropriate.
- If a table continues for more than one page, repeat the table heading with "(continued)" at the end, and repeat column headers for each page.

*For figures:*

- Two-dimensional figures work best.
- Label both axes of figures.

Here are some helpful tips for writing the results section: (a) Write with accuracy, brevity, and clarity. (b) Use past tense when referring to the results; use present tense when referring to tables and figures. (c) Do not interpret the reasons for the results achieved; such interpretation can be addressed in the discussion section. (d) If a lot of material is to be presented in the results section, subheadings may be helpful. (e) Do not include raw data or intermediate calculations in a research paper. (f) Do not present the same data more than once (e.g., in a table and then in a figure, or in a table or figure and then in the text). (g) Include results only from the study performed, not from other studies.

The results section is important because it describes the key results of a researcher's study. It should provide a clear and accurate description of the major findings in the form of text, tables, figures, or graphs.

*Bernard Choi and Anita Pak*

***See also*** Discussion Section; Methods Section

**Further Readings**

Branson, R. D. (2004). Anatomy of a research paper. *Respiratory Care, 49,* 1222–1228.

Choi, P. T. (2005). Statistics for the reader: What to ask before believing the results. *Canadian Journal of Anesthesia, 52,* R1–R5.

Hulley, S. B., Newman, T. B., & Cummings, S. R. (1988). The anatomy and physiology of research. In S. B. Hulley & S. R. Cummings (Eds.), *Designing clinical research* (pp. 1–11). Baltimore: William & Wilkins.

# Retrospective Study

In a retrospective study, in contrast to a prospective study, the outcome of interest has already occurred at the time the study is initiated. There are two types of retrospective study: a case–control study and a retrospective cohort study. A retrospective study design allows the investigator to formulate hypotheses about possible associations between an outcome and an exposure and to further investigate the potential relationships. However, a causal statement on this association usually should not be made from a retrospective study.

In conducting a retrospective study, an investigator typically uses administrative databases, medical records, or interviews with patients who are already known to have a disease or condition. Generally, a retrospective study is the method of choice for the study of a rare outcome, for a quick estimate of the effect of an exposure on an outcome, or for obtaining preliminary measures of association. After a brief discussion of the history of case–control and retrospective cohort studies, this entry discusses the use, advantages, and disadvantages of the retrospective study design.

## History

### *Case–Control Study*

As with other types of clinical investigation, the case–control study emerged from practices that originally belonged to the field of patient care. This form of disease investigation can be viewed as a combination of medical concepts (case definition, etiology, and a focus on the individual) and medical procedures (medical history taking, case series, and comparisons between the diseased and the healthy). The analytic form of the case–control study can be found in medical literature of the 19th century, but it did not appear to be viewed as a special or distinct methodology until the 20th century. The most fully developed investigation of

this type was Janet Lane-Claypon's 1926 study of breast cancer. This special field was crystallized in the years following World War II, when four case–control studies of smoking and lung cancer were published, and since then case–control study has been accepted as an approach to assessing disease etiology.

### *Retrospective Cohort Study*

Retrospective cohort studies have almost as long a history as the prospective studies. The first study was described by Wade Hampton Frost in 1933, based on assessment of tuberculosis risk in the black population in Tennessee. Interviews identified 556 persons from 132 families, which created 10,000 person-years of observation. In the presence of family contact, the attack rate of tuberculosis, standardized for age, was found to be about double that in the absence of such a contact (12.9 per 1,000 vs. 6.8 per 1,000). Although unable, because of the small number of people under study, to provide evidence of the importance of family contacts in the spread of tuberculosis, this study revealed the way records of past events could be used for the study of public health and is notable in particular for its clear description of the way person-years at risk can be calculated and its success in gaining the cooperation of almost an entire population.

Earlier studies of significance include the 1920s study by Bradford Hill of company records of nickel refinery workers and pensioners and the risk of cancers in the lung and nose, and the success in reducing this risk through changes in the refinery process, the study by R. E. W. Fisher on coal-gas workers' risk for lung cancer in 1920s, and many others.

## Use of the Study Design

### *Case–Control Study*

Case–control studies are retrospective, analytical, observational studies designed to determine the association between an exposure and outcome in which patients who already have a certain condition are compared with people who do not. Cases and controls are selected on the basis of the presence of the outcome condition, and exposure is assessed by looking back over time. It is very important in a case–control study that the cases be as similar to the controls as possible on all factors except the outcome of interest. Investigators then look to see what proportion in each group were exposed to the same risk factors.

These studies are commonly used for initial, inexpensive evaluation of risk factors and are particularly useful for rare conditions or for risk factors with long induction periods. Due to the nature of data sources, the results from case–control study are often less reliable than those from randomized controlled trials or cohort studies. In a case–control study, the statistical measure for the association is the *odds ratio*. A statistical relationship does not mean that one factor necessarily caused the outcome, however. Because of the potential for multiple forms of bias to occur in case–control studies, they provide relatively weak empirical evidence even when properly executed.

The case–control study is distinguished by six essential elements, each of which evolved separately in medical history.

1. The case definition: The disease entities are specific and are likely to have one or more specific causes.
2. An interest in disease etiology and prevention, in contrast to its prognosis or treatment.
3. A focus on individual, as opposed to group, etiologies.
4. Anamnesis, or history taking from patients, which permits the collapse of time past without enduring its slow passage until outcomes under study evolve.
5. Grouping individual cases together into series.
6. Making comparisons of the differences between groups in order to elicit average risk at the level of the individual.

### *Retrospective Cohort Study*

A retrospective cohort study allows the investigator to describe a population over time or obtain preliminary measures of association for the development of future studies and interventions. The exposure and outcome information in a cohort study are identified retrospectively by the use of

administrative data sets or through reviews of patient charts, interviews, and so on.

This type of study identifies a large population, studies patients who have a specific condition or have received a particular treatment over time, and compares them with another group that has not been affected by the condition or treatment being studied.

### *Case Reports and Case Series*

Case reports and case series are descriptive and exploratory, consisting of collection of reports on the treatment of individual patients or a report on a single patient. Because they are reports of cases and use no control groups with which to compare outcomes, they have no statistical validity. For example, a foot and ankle surgeon might describe the characteristics of an outcome for 100 consecutive patients with hammertoe who received a simple metacarpophalangeal joint release.

### Advantages and Disadvantages

One advantage of the retrospective study is that it is simple, quick, and inexpensive to conduct. It is also the only feasible method for studying very rare disorders or those with long lag between exposure and outcome. Other advantages are that fewer subjects are needed than for cross-sectional studies and that odds ratios can be calculated.

Some disadvantages include obtaining a single outcome and relying on recall or records to determine exposure status. In addition, retrospective studies must cope with confounders, risk of bias, and difficulty in selecting control groups. Furthermore, relative risk cannot be calculated, and there are no data on prevalence or incidence.

*Dongjiang Song*

***See also*** Association, Measures of; Cohort Design; Control Group; Interviewing; Observational Research; Odds Ratio; Prospective Study; Secondary Data Source

#### Further Readings

Doll, R. (2001). Cohort studies: History of the method: II. Retrospective cohort studies. *Sozial- und Präventivmedizin, 46*(3), 152–160.

Eyler, J. M. (2001). The changing assessment of John Snow's and William Farr's cholera studies. *Sozial- und Präventivmedizin, 46*(4), 225–232.

Hardy, A., & Magnello, M. E. (2002). Statistical methods in epidemiology: Karl Pearson, Ronald Ross, Major Greenwood and Austin Bradford Hill, 1900–1945. *Sozial- und Präventivmedizin, 47*(2), 80–89.

## Robust

Robust statistics represent an alternative approach to parameter estimation, differing from nonrobust statistics (sometimes called *classical statistics*) in the degree to which they are affected by violations of model assumptions. Whereas nonrobust statistics are greatly affected by small violations of their underlying assumptions, robust statistics are only slightly affected by such violations. Statisticians have focused primarily on designing statistics that are robust to violations of normality, due to both the frequency of nonnormality (e.g., via outliers) and its unwanted impact on commonly used statistics that assume normality (e.g., standard error of the mean). Nevertheless, robust statistics also exist that minimize the impact of violations other than nonnormality (e.g., heteroscedasticity).

### The Deleterious Effects of Relaxed Assumptions

In evaluating the robustness of any inferential statistic, one should consider both efficiency and bias. Efficiency, closely related to the concept of statistical power and Type II error (not rejecting a false null hypothesis), refers to the stability of a statistic over repeated sampling (i.e., the spread of its sampling distribution). Bias, closely related to the concept of Type I error (rejecting a true null hypothesis), refers to the accuracy of a statistic over repeated sampling (i.e., the difference between the mean of its sampling distribution and the estimated population parameter). When the distributional assumptions that underlie parametric statistics are met, robust statistics are designed to differ minimally from nonrobust statistics in either their efficacy or bias. When these assumptions are relaxed, however, robust statistics are

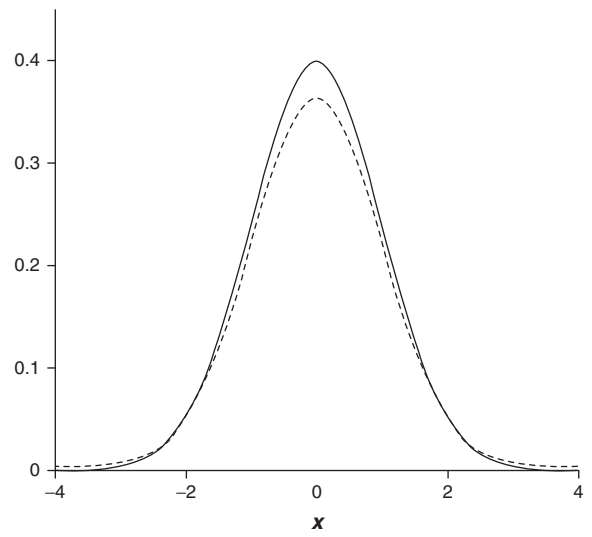



**Figure 1** A Normal Distribution[a] (black line), and a Mixed Normal Distribution[b] (dashed line)

*Source:* Adapted from Wilcox (1998).

*Notes: a.* $\mu = 0, s = 1$. b. Distribution 1 : $\mu = 0, s = 1$, weight = .9; Distribution 2: 1 : $\mu = 0$, $s = 10$, weight = .1.

designed to outperform nonrobust statistic in their efficacy, bias, or both.

For an example of how relatively minor nonnormality can greatly decrease efficiency when one relies on conventional, nonrobust statistics, consider sampling data from the gray distribution shown in Figure 1. This heavy-tailed distribution—designed to simulate a realistic sampling scenario in which one normally distributed population contaminates another ($\mu = 0$, $\sigma = 1$) with outliers—differs only slightly from the normal distribution, in black. Such nonnormality would almost surely go undetected by a researcher, even with large $n$ and when tested explicitly. And yet such nonnormality substantially reduces the efficiency of classical statistics (e.g., Student's *t, F* ratio) because of their reliance on a nonrobust estimate of population dispersion: sample variance. For instance, when sampling 25 subjects from the normal distribution shown in Figure 1 and an identical distribution one unit apart, a researcher has a 96% chance of correctly rejecting the null hypothesis via an independent-samples *t* test. If the same researcher using the sample sizes and statistics were to sample subjects from two of the heavy-tailed distributions shown in Figure 1—also spaced one unit apart—however, that researcher would have only a 28% chance of correctly rejecting the null. Modern robust statistics, on the other hand, possess high power with both normal and heavy-tailed distributions.

The above example is one of many realistic circumstances in which classical statistics can exhibit nonrobustness with respect to either bias or efficiency. Table 1 presents some other common circumstances. Advocates of robust statistics argue that assumptions such as normality are grounded in historical convention more than in empirical fact, observing that Gauss derived the normal curve to justify the use of sample means (not vice versa), as well as the frequently documented nonnormalities of real-world populations.

## Types of Robust Estimators

Today, the most widely used robust estimators of location are *M-estimators* and *trimmed means*. Whereas the former rely on the technique of *maximum likelihood estimation* (hence their name), the latter rely on the removal of a fixed proportion (commonly 20%) of a sample's smallest and largest observations. It is worth noting that although both of these sets of estimators characterize the "typical" value of a population based on sample data, neither are alternative estimates of the population mean: Instead, both estimate their respective population parameter (which, for symmetrical distributions, happens to equal the mean). Depending on the distribution, *M*-estimators may outperform (in terms of efficiency and bias) trimmed means or vice versa. For example, if outliers constitute more than 20% of the largest observations in a sample or are concentrated heavily in one tail of the distribution, *M*-estimators will likely outperform trimmed means.

To conduct a significance test with a robust estimator, one can rely on either parametric methods (e.g., *Yuen's test* for trimmed means) or nonparametric ones (usually *bootstrapping*).

## Why Not Transform Data or Remove Outliers?

Conventional wisdom asserts that researchers can combat the problematic effects of nonnormality without the use of robust statistics by transforming

**Table 1** Commonly Used Classical Statistics and General Conditions Under Which Studies Have Documented Their Nonrobustness

| *Statistic* | *Conditions of Efficiency Nonrobustness* | *Conditions of Bias Nonrobustness* |
|---|---|---|
| sample mean | asymmetrical distributions<br>heavy-tailed distributions | |
| sample variance | asymmetrical distributions<br>heavy-tailed distributions | |
| one-sample *t* | asymmetrical distributions<br>heavy-tailed distributions | asymmetrical distributions |
| two-sample *t* and ANOVA | heavy-tailed distributions<br>differences in symmetry<br>heteroscedasticity | differences in symmetry<br>heteroscedasticity |

*Source:* Adapted from Wilcox and Keselman (2003).

their data and removing outliers. Advocates of robust methods, however, point to several limitations in this approach. First, as the heavy-tailed distribution in Figure 1 makes clear, small perturbations of normality are difficult to detect with both visual and statistical analysis; tests of normality possess notoriously low power, and thus researchers often fail to detect—let alone correct for—nonnormality. Second, although curvilinear transformations can make distributions more normal, they cannot make distributions entirely normal; thus, normalizing transformations does not necessarily reduce the need for robust methods. Third, methods to detect outliers based on nonrobust statistics (e.g., sample mean and variance) often fail because of problems such as masking (in which one outlier masks the presence of another), and the deletion of outliers introduces unwanted dependence into the data that classical statistics fail to account for. Modern methods such as trimming and *M*-estimation account for the problematic effects of extreme data, and they do so in a robust manner that takes into account such dependence.

## Recommendations

As their pioneers have duly noted, robust statistics come with a cost of lower power under ideal circumstances of perfectly met model assumptions. For this reason, robust statistics have been described as a sort of insurance policy in which this small cost of lowered power protects researchers against the damaging effects of approximately met model assumptions. As with most situations in which two alternative analyses exist, however, the best policy is to blindly accept neither and instead conduct both. Thus, generally speaking, researchers are advised to conduct both classical and robust estimates, interpreting any meaningful differences in terms of either lowered power or relaxed assumptions.

*Samuel T. Moulton*

***See also*** Biased Estimator; Central Tendency, Measures of; Estimation; Influence Statistics; Influential Data Points; Nonparametric Statistics; Normal Distribution; Normality Assumption; Normalizing Data; Outlier; Precision; Trimmed Mean; Unbiased Estimator; Winsorize

### Further Readings

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics: The approach based on influence functions*. New York: Wiley.

Huber, P. J., & Ronchetti, E. M. (2009). *Robust statistics*. New York: Wiley.

Maronna, R. A., Martin, R. D., & Yohai, V. J. (2006). *Robust statistics: Theory and methods*. Chichester, UK: Wiley.

Tukey, J. W. (1960). A survey of sampling from contaminated distributions. In I. Okin (Ed.), *Contributions to probability and statistics* (pp. 445–485). Stanford, CA: Stanford University Press.

Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist, 53,* 300–314.

Wilcox, R. R., & Keselman, H. J. (2003). Modern robust data analysis methods: Measures of central tendency. *Psychological Methods, 8,* 254–274.

# Root Mean Square Error

The term *root mean square error* (*RMSE*) is the square root of *mean squared error* (*MSE*). *RMSE* measures the differences between values predicted by a hypothetical model and the observed values. In other words, it measures the quality of the fit between the actual data and the predicted model. *RMSE* is one of the most frequently used measures of the goodness of fit of generalized regression models.

## *RMSE* for Regression

In the application of regression models, unless the relationship or correlation is perfect, the predicted values are more or less different from the actual observations. These differences are prediction errors or residuals. These residuals are measured by the vertical distances between the actual values and the regression line. Large distances are indicative of large errors. However, for a given fitted regression line, the average or the sum of the residuals equals zero, as the overestimation of some scores can be canceled out by underestimations of other scores. Thus, a common practice in statistical work is to square the residuals to indicate the magnitude of absolute differences. Indeed, a primary goal in linear regression is to minimize the sum of the squared prediction errors to best model the relations among variables.

To acquire *RMSE,* one can square and average the individual prediction errors over the whole sample. The average of all the squared errors is the *MSE*. The *MSE,* in essence, is an average of the spread of the data around the regression line and reflects how big the "typical" prediction error is. Furthermore, the *MSE* can be "square rooted" to obtain *RMSE*. *RMSE* is used to represent prediction errors in the same units as the data, rather than in squared units.

## Mathematical Definition

Since *RMSE* is the square root of *MSE,* a thorough knowledge of *MSE* is important to an understanding of the mathematical definition and properties of *RMSE*.

### *MSE*

*MSE* is the mean of the overall squared prediction errors. It takes into account the bias, or the tendency of the estimator to overestimate or underestimate the actual values, and the variability of the estimator, or the standard error.

Suppose that $\hat{\theta}$ is an estimate for a population parameter θ. The *MSE* of an estimator $\hat{\theta}$ is the expected value of $(\hat{\theta} - \theta)^2$, which is a function of the variance and bias of the estimator.

$$MSE(\hat{\theta}) = E([\hat{\theta} - \theta]^2) \text{ and}$$

$$MSE(\hat{\theta}) = V(\hat{\theta}) + (B[\hat{\theta}])^2,$$

where $V(\hat{\theta})$ denotes the variance of $\hat{\theta}$ and $B(\hat{\theta})$ denotes the bias of the estimator $\hat{\theta}$.

An *MSE* of zero means that the estimator $\hat{\theta}$ predicts observations of the parameter θ perfectly. Different values of *MSE* can be compared to determine how well different models explain a given data set. The smaller the *MSE* is, the closer the fit is to the data.

### *RMSE*

*RMSE* is the average vertical distance of the actual data points from the fitted line. Mathematically, *RMSE* is the square root of *MSE*.

$$RMSE(\hat{\theta}) = MSE(\hat{\theta})^{1/2} = (E[(\hat{\theta} - \theta)^2])^{1/2}$$

For an unbiased estimator, *RMSE* is equivalent to the standard error of the estimate, and it can be calculated using the formula

$$s = \frac{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2}}{n - 1},$$

where $n$ denotes the size of the sample or the number of observations; $X_i$ represents individual values, and $\overline{x}$ represents the sample mean.

In many statistical procedures, such as analysis of variance and linear regression, the *RMSE* values are used to determine the statistical significance of the variables or factors under study. *RMSE* is also used in regression models to determine how many

predictors to include in a model for a particular sample.

### Properties

As the square root of a variance (*MSE*), *RMSE* can be interpreted as the standard deviation of the unexplained variance. Thus, *RMSE* is always above zero. Its minimum occurs only when the estimate is indefinitely close to the real data. In other words, *RMSE* is an indicator of the fit between an estimate and the real data points. Smaller *RMSE* reflects greater accuracy.

However, there is no absolute criterion for an ideal value for *RMSE,* because it depends on the scales of the measured variables and the size of the sample. *RMSE* can be compared only between models whose errors are measured in the same units. While a particular value of *RMSE* greater than zero is not meaningful in and of itself, its magnitude relative to other *RMSE* values can be used to compare models. A model with a smaller *RMSE* is often preferred.

As the squaring process gives disproportionate weight to large errors, the *RMSE* is very sensitive to the occasional large error. When many extreme values are present, a model is likely to get large values on the *RMSE*.

### Relations to Other Fit Measures

As *RMSE* has the same units as the predicted quantity, it is more easily interpretable than *MSE* is. In addition, *RMSE* is associated with the $R^2$: Large *RMSE* corresponds to low $R^2$. To be more specific, *RMSE* can be computed by the standard deviation of the sample and $r^2$. However, *RMSE* is a better measure of goodness of fit than is a correlation coefficient because *RMSE* is interpretable in terms of measurement units of the assessed variables. *RMSE* is also more advantageous than the mean absolute error in detecting extremely large errors, because the mean absolute error is less sensitive to extreme values.

*Yibing Li*

***See also*** Standard Error of Estimate; Standard Error of Measurement; Standard Error of the Mean

**Further Readings**

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2002). *Applied multiple regression: Correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Degroot, M. H. (1991). *Probability and statistics* (3rd ed.). Reading, MA: Addison-Wesley.

Kenney, J. F., & Keeping, E. S. (1962). Root mean square. In J. F. Kenney & E. S. Keeping (Eds.), *Mathematics of statistics* (3rd ed., pp. 59–60). Princeton, NJ: Van Nostrand.

Lehmann, E. L., & Casella, G. (1998). *Theory of point estimation* (2nd ed.). New York: Springer.

Wackerly, D. D., Mendenhall, W., & Scheaffer, R. L. (2002). *Mathematical statistics with applications* (6th ed.). Pacific Grove, CA: Duxbury.

## Rosenthal Effect

A general characteristic of human nature is that people tend to judge themselves, especially their competence and worth, based on the perception of others. The term *Rosenthal effect* refers to this internalization of expectations from a perceived authority figure by the recipient. Four terms are used to describe this same phenomenon, generally interchangeably: *Pygmalion effect, Rosenthal effect, self-fulfilling prophecy,* and *expectancy effect*. Another associated term—*Hawthorne effect*—underscores the complete understanding of the Rosenthal effect. In this entry, these terms are further examined, and implications for research and practical applications, particularly within the education system, are examined for the Rosenthal effect. It is imperative that researchers obtain a clear understanding of the Rosenthal effect because the impact of another person's expectations of recipients can generate very powerful effects on them and their resultant behaviors. As a result, researchers may wish either to research this phenomenon further or guard against its impact when collecting data.

### History

The term *Pygmalion* effect first came into existence with the Greek mythological tale of

Pygmalion and Galatea. Pygmalion was a sculptor who wished to create the perfect woman, whom he proceeded to carve from ivory. His creation was the epitome of his idea of womanhood, and he wished with all his heart for her to be a real woman. He prayed to the goddess Aphrodite and entreated her to allow Galatea to become flesh and blood. The goddess granted his wish, and Pygmalion and Galatea became husband and wife.

In 1916, George Bernard Shaw wrote the play *Pygmalion,* in which the lead character, Professor Higgins, transformed a common flower girl into a lady, so much so that he could pass her off as a duchess. The flower girl internalized the fact that the professor was convinced that she could metamorphose, and this boosted her own confidence in her ability to successfully transcend her meager and lowly beginnings to become—the desire of her own heart—a lady. It was from these stories that the term *Pygmalion effect* was coined. Specifically, Shaw's plot demonstrated a real-world application of the expectancy effect. However, both of these stories reflect one basic premise: that if a person believes that something can be achieved, then it really can be achieved!

## Real-World Application

The Rosenthal effect was named in honor of Robert Rosenthal, who studied the expectancy phenomenon extensively, originally with a student sample. To best illustrate the power wielded by the Pygmalion effect, Rosenthal and Lenore Jacobson conducted research at an elementary educational institution. They were curious about what would happen to the expectations of both teachers and students if students were unsuspectingly and randomly chosen and touted as having high potential. A deception involved the students' having to take a supposedly highly credible test that would make the predictions about the student's potential. The results of the deception on subsequent teacher expectations—which in turn influenced the levels of student confidence in their own potential—confirmed the Pygmalion/Rosenthal effect. Once the high-potential students were identified by the "test," the researchers found that when positive expectations were evident from teachers, the students were motivated to strive toward the goal due to renewed or increased self-confidence. As a result, success was usually and significantly apparent. However, the converse also proved to be true: When teacher expectations about students were negative, the students become convinced of their incompetence, and failure was the common result.

## When the Rosenthal Effect Meets the Hawthorne Effect

Associated with the Rosenthal effect or the self-fulfilling prophecy is the Hawthorne effect. The research carried out by Rosenthal and Jacobson can serve to illustrate the relationship. For example, the Hawthorne, or *reactivity,* effect can result in improvement in task performance because more attention is provided to a group of people. From the example, the teachers may have, either consciously or unconsciously, provided more attention and more praise to the students who were identified as having high potential. As a result, the students bloomed under the additional attention and confirmed the positive teacher expectations through what is termed the *self-fulfilling prophecy* or *Rosenthal effect*. These results have been confirmed by other researchers.

Research into the Rosenthal effect and self-fulfilling prophecy indicated that the earlier that teachers were informed about their students' potential, the greater would be the advantages. Further research suggested that it was necessary for the teacher who initially expected the students to perform well to continue to influence the students. If the original teacher was replaced as the students progressed through higher grades, then the expectancy effect tended to decrease markedly. If, however, the students continued to be exposed to the original teacher, then the expectancy effect was maintained. Research has also shown that younger children are more susceptible to the Rosenthal effect, but older children tend to be able to retain the effect for a longer time period. As the example indicates, the Rosenthal effect has been identified within the classroom with regards to performance on tests. However, it has also been found in response to teacher attitudes about minority students and gender bias. Areas of study on expectancy effects in therapy have also been included.

## The Rosenthal Effect on Animals

Rosenthal conducted studies on *experimenter expectancy effect,* which encapsulates the idea that experimenters tend to confirm the existence of what they expect to find. In one such experiment with animals, Rosenthal identified two sets of rats: one set was said to have been smart enough to learn the maze while the other group was downplayed as the exact opposite. In reality, both groups of rats were from the same set of offspring. Once it was indicated that the groups were different and the specific nature of that difference was outlined, the experimenter in the study would proceed to confirm that the "smarter" rats performed better at the maze task than did the other group.

## Advantages of Understanding the Rosenthal Effect

The hundreds of studies conducted on the Rosenthal effect have not only added to the current literature but have also provided a pathway for practical application. The work of Rosenthal and his colleagues has in fact triggered other researchers to try to realign effects of negative expectations that confront certain groups. This shift suggests that investigation into the expectancy phenomenon and its associated implications can provide the springboard for deliberate and effective strategies to convert that which is negative (e.g., stereotypes) into a positive outcome.

Although much of the work on the Rosenthal effect has been conducted by Rosenthal and his colleagues, other researchers have also found that teacher expectations do in fact enhance performance despite the actual ability of the students in question. Examples in support of these findings range from the progress of students in Head Start programs to that of intellectually challenged children. However, some researchers have been severely critical of the methodology employed by Rosenthal. Over the years, Rosenthal addressed these methodological issues, with the result that his current studies are reflective of greater scientific rigor.

## Limitations on Research and Possible Solutions

Unless the Rosenthal effect and its correlates are being investigated specifically, the results of other types of research may be confounded by the expectancy effect. Even within a laboratory environment, there are many factors besides the independent variable that can impact the results. For example, if the Hawthorne effect is added to the experimenter expectancy effect, two methodological problems occur that may lower the internal validity of the study. First, people tend to do better when they realize that they are under observation; the added attention can cause them to perform better even if conditions are made worse than usual. Second, experimenters may either consciously or unconsciously look only for evidence to confirm their hypothesis. In order to surmount these problems, two strategies may be employed. First, a control or placebo group may be formed. If the results of this group are almost the same as or as good as the group in which the independent variable is being examined, then the experimental methodology needs to be revised. Second, to minimize the impact of experimenter expectancy effects, a double-blind procedure may be used in which both the participants and the experimenters are unaware of the characteristics of the groups under investigation. For example, if three groups have to perform a task but the experimenter does not know which group is smart, which group is obtuse, and which group forms the control for the experiment, then this situation minimizes the experimenter expectancy effect. It also follows that if participants do not know to which group they belong, the Rosenthal effect and self-fulfilling prophecy may be avoided.

Even if experimenters use standardized instructions, nonverbal and totally unconscious cues may be picked up by the participants. One of the most famous cases of unconscious cuing is that of Clever Hans. This horse, trained by William Von Osten, was able to answer many questions involving the alphabet, mathematical calculations, and names of people by pawing the ground. No evidence of cheating was ever uncovered, and researchers were baffled by the horse's ability. To fully test him, he was placed in a tent, devoid of spectators and consequent distractions, with a panel of scientists who provided questions on all

facets of knowledge that Hans was supposed to know. What they discovered was that Hans was actually responding to unconscious cues provided by the various questioners. For example, when the questioner's face relaxed, Hans stopped pawing the ground as this became a cue that the answer was correct. If the questioner's face was tense, then the answer had not been obtained as yet. If Hans was blindfolded, he was totally inaccurate in his responses. It was concluded that the questioners were unconsciously cuing the horse. Thus, experimenter expectancy bias was certainly evident in animals and in humans.

Therefore, another strategy that may lessen the impact of experimenter expectancy effects via unconscious cuing is the use of audiovisual technology to record the instructions. The use of a prerecording maintains the standardization and avoids any extra cues that do not exist in the original recording. Furthermore, the participants may be able to record their answers on audiovisual technology in order to preserve the objectivity of the study and avoid having any cues from a tester influence the answers of the participant. For example, not all testers can maintain expressionless features, especially if the participant is doing extremely well or extremely poorly. The participant's recording answers can help increase the objectivity of the task at hand.

*Indeira Persaud*

***See also*** Experimenter Expectancy Effect; Hawthorne Effect; Internal Validity; Reactive Arrangements

**Further Readings**

Alvidrez, J., & Weinstein, R. S. (1999). Early teacher perceptions and later student academic achievement. *Journal of Educational Psychology, 91,* 731–746.

Aronson, J., Fried, C. B., & Good, C. (2002). Reducing the effect of stereotype threat on African American college students by shaping theories of intelligence. *Journal of Experimental Social Psychology, 38,* 113–125.

Babbie, E. (2004). *The practice of social research* (10th ed.). South Melbourne, Victoria, Australia: Thomson Wadsworth.

Chaiken, A., Sigler, E., & Derlaga, V. (1974). Non-verbal mediators of teacher expectancy effects. *Journal of Personality & Social Psychology, 30,* 144–149.

Kaplan, R. M., & Saccuzzo, D. P. (1997). *Psychological testing: Principles, applications and testing* (4th ed.). Pacific Grove, CA: Brooks/Cole.

Literature Network. (2000). *Pygmalion* [Online]. Retrieved April 6, 2009, from http://www.online-literature.com/george_bernard_shaw/pygmalion

Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the classroom: Teacher expectation and student intellectual development.* New York: Holt, Rinehart & Winston.

Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioural research: Methods and data analysis* (2nd ed.). New York: McGraw-Hill.

# RUBRICS

Rubrics are descriptive scoring systems that allow observers to assign a numeric value to a piece of work or a performance. While rubrics or scoring guides are most often associated with the field of measurement, specifically performance assessment, the use of such assessment tools can be applied more generally to any research endeavor in which one intends to apply a rating or a score to a complex task, product, or process. Rubrics provide a set of guidelines or criteria that allow researchers or evaluators to describe a set of characteristics that represent an underlying continuum of performance. The use of order categories that represent the different levels associated with a given performance or product allows for subjective judgments to become more objective in nature. This entry focuses on developing and assessing the quality of rubrics.

## Rubric Development

Rubrics may be characterized as general or task-specific and may use an analytic or holistic scoring system. The purpose of the assessment determines the type of rubric most appropriate for use. For example, if the purpose is to gather information regarding specific components of performance, then a task-specific rubric may be the most appropriate tool. On the other hand, if the purpose is to examine a broader range of tasks, then the evaluator may select a more general rubric for scoring. In much the same way, the scoring system itself can be either analytic or holistic in nature. When

a product or performance comprises a variety of criteria on which judgment is made and each of these criteria is set along a continuum for scoring, the rubric is considered to be analytic. In cases in which only one score is assigned without the emphasis on specific criteria associated with the performance or product, the rubric is considered to be holistic. An example of a commonly used holistic scoring system is the assignment of grades, such as A, B, C, D, and F. Again, the format of the rubric is determined by the purpose of the evaluation.

General guidelines have been established in an effort to assist in the development of scoring rubrics. The first step is to identify the objective of the evaluation. Once this has been determined, the evaluator should identify a series of observable attributes and then thoroughly describe characteristics associated with each. Once the attributes have been identified and the characteristics determined, narratives are written along the continuum (either holistic or analytic), and score points are assigned at each interval. Prior to full implementation of the scoring system, anchors of performance are gathered to ensure that all levels of performance have been identified. Finally, rubrics are revised as needed. The key to the development of any rubric is clearly identifying the goals and objectives to be evaluated. Clear objectives for scoring the product or performance will help determine the format as well as the levels of performance to be scored.

Rubrics can be either too task-specific or too general, having the potential to make scoring difficult. Because rubrics are a measurement tool, attention must be paid to the reliability and validity of the scoring system. Development of a quality rubric is based primarily on the ability of a judge or rater to apply the performance criteria in an objective and consistent manner, ensuring that the scoring criteria adequately measure the objectives of the assessment. The descriptive characteristics of each score point along the scoring continuum may need to be reexamined and sometimes rewritten so that raters can accurately and consistently identify the level of performance based on the descriptors provided. Consistency within and across raters is crucial to the reliability of the rubric. In addition, the degree to which the content or the attributes of performance are aligned with stated goals and objectives should be examined. In some cases, comparisons with existing measures may be useful in helping one determine the quality of the rubric. As with any measurement tool, the overall goal is to collect validity evidence that supports the rubric's ability to accurately measure the underlying psychological construct for which it was intended.

## Rubric Quality

When assessing the quality of a rubric, one might consider four components: (1) purpose, (2) criteria, (3) scoring, and (4) organization of the instrument. The purpose of the assessment should be clearly stated and objectives identified. Evaluators must also determine whether the use of a rubric is the best method of measurement for any given situation. Once the purpose or objective(s) are clearly delineated, one must determine whether the performance criteria are clearly written. Clarity and detail increase the objectivity of the rubric for evaluators. The degrees of performance must also be precisely measured through an accurate scoring system that represents a continuum of performance represented by significant differences among score points. Finally, the overall organizational structure of the rubric must be examined to be sure that spelling, grammar, and physical layout are appropriate and user friendly.

Although rubrics may be developed for any situation in which one wishes to judge a performance or product, they have become most popular among educators, especially classroom teachers. The use of rubrics in educational settings has the potential to influence learning by providing formative feedback to learners on their progress as well as providing feedback to instructors on teaching.

*Vicki L. Schmitt*

***See also*** Criterion Variable; Standardization

### Further Readings

Goodrich, H. (1996). Understanding rubrics. *Educational Leadership, 54*(3), 14–17.

Mertler, C. A. (2001). Designing scoring rubrics for your classroom. *Practical Assessment, Research and Evaluation, 7*(25). Retrieved March 1, 2010, from http://pareonline.net/getvn.asp?v = 7&n = 25

Moskal, B. M. (2000). Scoring rubrics: What, when and how? *Practical Assessment, Research and Evaluation, 7*(3). Retrieved March 1, 2010, from http://pareonline.net/getvn.asp?v = 7&n = 3

Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research and Evaluation, 7*(10). Retrieved March 1, 2010, from http://pareonline.net/getvn.asp?v = 7&n = 10

Popham, W. J. (1997). What's wrong and what's right with rubrics. *Educational Leadership, 55*(2), 72–76.

Tierney, R., & Simon, M. (2004). What's still wrong with rubrics: Focusing on the consistency of performance criteria across scale levels. *Practical Assessment, Research & Evaluation, 9*(2). Retrieved March 1, 2010, from http://pareonline.net/getvn.asp?v = 9&n = 2

# S

## Sample

Samples are used in much of the research conducted in both the social and natural/physical sciences. Even in the fields of history, music, and English, samples are often used in empirical investigations. Because the quality of the sample(s) used in a study has such an important bearing on the trustworthiness of a researcher's conclusions, here the following six critical aspects are considered:

1. The definition of *sample*
2. Why researchers use samples
3. How sample data are used
4. What kinds of samples researchers use
5. Why the concept of "sampling error" is important
6. The misconceptions people have about samples

### Definition and Examples

A sample is a subset of a population. The population can be a set of people, animals, or things. For example, the population might be made up of all children who, on a given day, are enrolled in a particular elementary school; the dogs, during a given month, that are treated at a particular veterinary hospital; or the boats, on a given evening, that are docked at a particular marina. With these entities as the possible populations, a sample would be a *subset* of school children, a *subset* of dogs, or a *subset* of boats, respectively. The notion of *subset* implies that a sample must be smaller than the population to which it is connected. If $n$ represents the size of the sample and $N$ represents the size of the population, then it is mandatory that $n < N$. If the sample and the population are identical in quantity, then only the population exists.

### Reasons Why Researchers Use Samples

Most researchers use samples in their studies because they do not have the resources—time, money, or access—to measure or observe all members of the population of interest. If a researcher wants to describe the eating behavior of the dogs observed at a particular veterinary hospital, the researcher's available time and money would likely prohibit the in-home observation of every dog. However, observing a subset of those dogs might be feasible given the study's budget and duration. Here, as in many other situations, the researcher chooses "what's possible" (even though it is not best) instead of "what's best" (because it is not possible).

In certain studies, researchers use samples because they have no choice. In laboratory and field experiments, and in clinical trials, a group of people, animals, or things is given some form of treatment. Posttreatment measurements are used as a basis for comparing the treated group against a control group, a placebo group, or a different treatment group. Or the treatment group's

posttreatment scores are sometimes compared against the same group's pretreatment scores. In such investigations, the treatment group typically is considered to be a sample of a larger group that, in the future, might also receive the treatment. For obvious reasons, it is not possible to observe or measure the entire population.

## Typical Use of Sample Data

After the members of a sample are identified and measured, researchers normally summarize the sample data and then use what they know about the sample to make an educated guess (i.e., a statistical inference) about the population. For example, if the population of interest is defined as the freshmen who enter a particular university in the fall of a given year, a sample of those freshmen might be interviewed to find out what they worry about. If these interviews reveal that 60% of the students in the sample say they worry about "doing well academically," this information concerning the sample—referred to as the *statistic*—would be used to make a guess about the percentage of students in the freshman class who worry about grades—the *parameter.*

Typically, the sample-based inference involves one of two kinds of statistical thinking and analysis. One is *estimation;* the other is *hypothesis testing.* In estimation, the sample statistic can be used by itself as an estimate of the population parameter, or a confidence interval can be created to generate a "ball-park estimate." In hypothesis testing, the sample data are used to determine whether a possible value of the population parameter, which is contained in the null hypothesis, can be rejected.

## Types of Samples

Many different kinds of samples are used in research studies. The various types of samples differ from one another mainly in terms of the procedure used to determine how members of the sample are selected. The two main categories of samples are probability samples and nonprobability samples.

### *Probability Sampling Techniques*

A probability sample is a subset of a population in which each member of the population has a precise probability of being included in the sample. For example, in a Bingo game where balls numbered 1 through 75 are blindly selected until a contestant has a winning card, the probability is $\frac{1}{75}$, or $.013\overline{3}$, that any given ball (e.g., the ball with the number 34) will be the first ball selected. Although there are many types of probability samples, the following four are frequently used in research investigations: simple random samples, systematic random samples, stratified random samples, and cluster samples.

#### Simple Random Samples

With simple random samples, each member of the population must have an equal probability of being included in the sample. It follows that each possible sample of size $n$ has an equal probability of being selected from the population. After a list of the population's members (i.e., the sampling frame) is created, a random number generator, a random numbers table, or some other unbiased method is used to generate the sample. The method used can involve *sampling without replacement* or *sampling with replacement.* With the former method, members of the sampling frame are eliminated from the list once they are selected for inclusion in the sample; with the latter method, the sampling frame remains intact and thus makes it possible for members of the population to appear two (or more) times in the sample.

#### Systematic Random Samples

Systematic random sampling is another way to select a random sample from a population. Four steps are involved. First, the researcher must specify the number of members in the population of interest ($N$). Second, the desired size of the sample ($n$) must be specified. Third, the researcher divides $N$ by $n$—with this result rounded, if necessary, to the nearest integer—to establish the *sampling interval* ($k$). Finally, after establishing the values of $N$, $n$, and $k$, the sample is created in $n$ stages. First, a member of the population is randomly selected from among positions 1 through $k$ in the sampling frame. Next, $k$ is added to that initial "starting

spot" so as to identify the second member of the sample. The third member of the sample is identified by adding *k* to that value, and so on, until the specified sample size is reached.

#### Stratified Random Samples

If the population contains two or more mutually exclusive categories (i.e., strata), a stratified random sample can be created. To do this, simple random sampling is used within each stratum to identify members of the sample. Often, the size of each stratum in the sample is made to be proportionate to corresponding stratum in the population. For example, if a population is made up of 70% men and 30% women, the sample might be set up to contain these same sex proportions. Researchers sometimes *oversample* one or more of the population's strata; this is done if it is known that the members of certain strata are less likely to return surveys, less likely to remain in a study until its completion, and so on.

#### Cluster Samples

Sometimes, naturally occurring samples exist. For example, students in a classroom within a university represent a cluster of potential participants for a researcher interested in evaluating the effects of a new curriculum. Or the members comprising a population might be so spread out geographically that simple random sampling could be difficult or even impossible. A researcher using cluster sampling first creates a list of all possible clusters within the population of interest (e.g., counties in a state or classes within a school). Then, a random sample of these clusters is chosen. Finally, the researcher either selects all subjects from these selected clusters (*single-stage cluster sampling*) or selects a random sample of subjects from within them (*two-stage cluster sampling*).

### *Nonprobability Sampling Techniques*

Instead of starting with an existing, tangible population and then extracting a sample from a sampling frame, researchers often begin with a sample and then create an imaginary population that fits the sample. Such samples are called *nonprobability sample*s, and they differ from probability samples in two main respects, as follows: (1) the temporal sequence in which the sample and population come into existence and (2) the nonuse of random sampling procedures.

With both probability and nonprobability samples, the typical goal is to generalize information gleaned from the sample to the relevant population. As indicated earlier in this discussion, such generalizations often involve estimation and/or hypothesis testing. These two kinds of statistical procedures can be used with nonprobability samples just as easily as they can be used with probability samples. The main difference is that the statistical inference based on a nonprobability sample is aimed at a population that is abstract rather than concrete.

The four main types of nonprobability samples are convenience samples, quota samples, snowball samples, and purposive samples.

#### Convenience Samples

Sometimes researchers choose their samples based on who or what is available at the time an investigation is undertaken. Such samples are called convenience samples. (Convenience samples are also called *haphazard* or *accidental* samples.) For instance, many researchers gather data from college students (especially freshmen) because they are readily available and willing to participate in research. Or suppose a researcher goes to a local nursing home and measures the mobility of all residents. If the resulting data from this study are used to generate confidence intervals or to evaluate stated null hypotheses, then those data can be said to come from a convenience sample.

#### Quota Samples

Quota samples are similar to stratified random samples; however, quota sampling does not use random selection. For example, suppose that a researcher wishes to correlate people's gender and hand preference in a city where there are as many males as females. Suppose also that this researcher stands at a busy street corner in a large city and asks this simple question to the first 50 males and the first 50 females who pass by: Are you right-handed, left-handed, or ambidextrous? The resulting sample of 100 people in this study is a quota sample. Once the predetermined number of male or female slots is filled, no more

people of that gender would be queried about their handedness.

#### Snowball Samples

Snowball sampling is used by researchers who need to identify people who, for one or more reasons, are difficult to find. For example, suppose a researcher wants to interview natives of Guam living in Florida. There most likely is no directory of such individuals from which a sample could be drawn. Therefore, the researcher might ask a few Guam-born Floridians to identify others they know who match the study's criteria. Those newly identified individuals might be asked to identify additional people living in Florida who came from Guam. Although snowball sampling, in certain studies, can generate an impressively large sample, the valid population is abstract in nature and can be defined only after the sample is formed.

#### Purposive Samples

In certain studies, the nature of the research questions necessitates that certain criteria be used to determine who or what goes into the sample. For example, suppose a market researcher wants to explore the marketability of a new twin stroller. Also suppose that this researcher would like to tap the opinions of women between the ages of 20 and 30 who have twins younger than 3 years. If data are collected from a group of individuals who possess these specific characteristics (mothers in their 20s with twins under the age of 3), then that group of individuals is a purposive sample. The relevant population to which valid generalizations can be made is abstract and cannot be conceptualized until we know where the sample was located geographically, what inducements (if any) were offered for participating in the investigation, what experience (if any) the mothers had with strollers, and so on.

## The Concept of "Sampling Error"

The data on a variable of interest gathered from a sample usually are summarized into a condensed, single-number representation (e.g., mean) of the sample called the sample *statistic*. Using estimation or hypothesis testing, the observed data from the sample are used to make an inferential statement about the population's unknown standing on the targeted variable, which is the population *parameter*.

Sampling error exists if the sample statistic differs in size from the population parameter. There usually will be a discrepancy between the sample statistic and the population parameter, even with randomly selected probability samples. The precise amount of sampling error is usually unknown because the magnitude of the population parameter is not known. However, the data from a sample can be used to estimate how probable it is that the sample statistic lies any given distance away from the population parameter. Such probabilities constitute the foundation of the statistical procedures that lead to confidence intervals or decisions to reject or fail to reject null hypotheses.

The likely size of sampling error can be determined by using sample data to estimate the *standard error* of the sample statistic. For example, if a study's focus is on the mean, then the sample data can be used to estimate the standard error of the mean $SE_M$ via the formula

$$SE_M = \frac{s}{\sqrt{n}},$$

where $s$ is the sample-based estimate of $\sigma$ and $n$ is the size of the sample. The estimated standard error provides a scientific guess as to amount of variability that would exist if many samples of the same size were drawn from the same population. Just as a standard deviation measures the amount of spread in a set of numbers, the estimated standard error indicates the expected amount of dispersion in the values of the sample statistic.

## Misconceptions About Samples

### *A Random Sample Will Be Identical to the Population, Only Smaller*

A common misconception about random sampling is that a random sample drawn from a population will have the exact same numerical characteristics as that population. In other words, many people think that a sample will have the same mean, median, standard deviation, and so on, as the population from which it was drawn. This is a misconception because sampling error is

likely to be present rather than be absent when a sample is extracted from a population.

Suppose a simple random sample of 100 people is taken from a large population containing equal numbers of males and females. It is possible, of course, for the sample to have as many males as females. However, the binomial distribution reveals that such an outcome has less than an 8% chance of occurring. Change the variable of interest from something that is dichotomous (such as gender) to something that is continuous (such as height or weight) and it is far less likely that the sample will be just like the population.

### *The Distribution of the Sample Statistic Has the Same Form as the Population's Distribution*

Another common misconception about sampling concerns distributional shape. Many people think that the distributional form of a sample statistic, for example, the mean, will be identical to the shape of the population. People who believe this expect the sampling distribution of the mean to be negatively skewed if the population is negatively skewed, with both distributions having the same degree of skewness. This is not the case.

The Central Limit Theorem states that the sampling distribution of $\overline{X}$ tends toward normality as $n$ increases, *regardless of the shape of the population.* Moreover, simulation studies have shown that the sampling distribution of the mean is fairly normal even if $n$ is small and even if the population is far from normal. For example, if the sample size is 10 and if the population is rectangular, it is striking how close to normal the sampling distribution of $\overline{X}$ is. Increase $n$ to 25, and the sampling distribution appears exceedingly bell shaped (with only slight deviations from the skewness and kurtosis indices of the normal curve).

### *Large Populations Require Large Samples*

Intuition leads many people to believe that a sample's quality is related to the relative size of $n$ to $N$. According to this line of reasoning, a sample needs to be larger if it is being drawn from a larger population. In reality, it is the absolute size of $n$ that most influences the quality of a sample, not the relative size of $n$ to $N$. This becomes apparent if one examines the formula for the standard error of the statistic. When a finite population exists, the size of $N$ plays a meaningful role only when a tiny sample is drawn from a small population. When $n$ is not small, the size of the standard error is determined almost totally by $n$; in this situation, it matters little whether $N$ is slightly or greatly larger than $n$. This is why political polls can end up being so accurate even though they are based on data from only a tiny fraction of the sampled population.

*Schuyler W. Huck, Amy S. Beavers, and Shelley Esquivel*

***See also*** Central Limit Theorem; Cluster Sampling; Confidence Intervals; Convenience Sampling; Estimation; Parameters; Probability Sampling; Quota Sampling; Random Sampling; Sample Size; Sampling; Sampling Error; Statistic; Systematic Sampling

**Further Readings**

Benedetto, J. J., & Ferreira, P. J. S. G. (Eds.). (2001). *Modern sampling theory: Mathematics and applications*. Boston: Birkhäuser.

Deming, W. E. (1984). *Some theory of sampling*. Mineola, NY: Dover.

Hedayat, A. S., & Sinha, B. K. (1991). *Design and inference in finite population sampling*. New York: Wiley.

Henry, G. T. (1990). *Practical sampling*. Newbury Park, CA: Sage.

Lohr, S. L. (1999). *Sampling: Design and analysis*. Pacific Grove, CA: Duxbury Press.

Rao, P. S. R. S. (2000). *Sampling methodologies*. Boca Raton, FL: CRC Press.

Sampath, S. (2001). *Sampling theory and methods*. Boca Raton, FL: CRC Press.

Shaeffer, R. L. (2006). *Elementary survey sampling*. Belmont, CA: Thomson Brooks/Cole.

Thompson, S. K. (2002). *Sampling*. New York: Wiley.

## Sample Size

Sample size refers to the number of subjects in a study. If there is only one sample, then the sample size is designated with the letter "$N$." If there are samples of multiple populations, then each sample size is designated with the letter "$n$." When there are multiple population samples, then the

total sample size of all samples combined is designated by the letter "*N*."

A study's sample size, or the number of participants or subjects to include in a study, is a crucial aspect of an experimental design. Running a study with too small of a sample runs numerous risks including not accurately reflecting the population a sample was drawn from, failing to find a real effect because of inadequate statistical power, and finding apparent effects that cannot be replicated in subsequent experiments. However, using more subjects than necessary is a costly drain on resources that slows completion of studies. Furthermore, if an experimental manipulation might pose some risk or cause discomfort to subjects, it is also ethically preferable to use the minimum sample size necessary. This entry focuses on the factors that determine necessary sample size.

## Magnitude of Expected Effect

In general, when possible it is preferable to design an experiment looking for large effects that can be more easily detected with relatively small sample sizes. However, in many cases, small to modest effects can still be very important. For instance, important psychological processes might be associated with relatively subtle changes in detectible biological measures, such as the relatively minor changes in blood oxygen level dependent signaling (BOLD) that corresponds with neural activity. Additionally, treatments that produce relatively modest clinical improvements could still significantly benefit and improve the quality of life of afflicted individuals. Likewise, an even slightly more accurate diagnosis process could save countless lives. For studies looking to detect relatively small to modest effects, larger sample sizes will be needed.

## Variability

The variability of data is a crucial factor for estimating what sample size is needed. The sample sizes needed in descriptive studies are dependent on the variability of measures of interests in the population at large. If the measures of interest are narrowly distributed in the population, then smaller sample sizes might be sufficient to predict these measures accurately. Alternatively, if these measures are broadly distributed in the population, then larger sample sizes are needed to predict these measures accurately.

For example, suppose a group of forestry students wished to determine the average tree height on a Christmas tree farm and in an adjacent forest. All the trees on this Christmas tree farm are 4-year-old Douglas firs, whereas the trees in the forest are of various ages and species. The students could likely measure relatively few trees in the Christmas tree farm and have an accurate idea of the average tree height, whereas they would likely have to measure many more trees in the forest to determine the average tree height there.

In experimental studies, the more variable data are across subjects, the more subjects will be needed to detect a given effect. One means of reducing variability across subjects and thereby reducing the sample size required to detect an effect is to use a within-subject design. Within-subject designs, or repeated testing on the same subjects across the different phases of the experiment, reduces variability across subjects by allowing each subject to serve as his or her own control. Care must be taken to control for possible carry-over effects from prior testing that might influence later measures.

For example, let us suppose a researcher wishes to test the effects of a novel drug on performance on a memory task. Under baseline conditions, there is a fair amount of variability in performance across subjects, but each subject scores about the same each time he or she performs the task. Using a within-subjects design and comparing each subject under placebo and drug conditions helps control for variability between subjects and allows the researcher to use a smaller sample size than would be necessary in a between-subject design with separate placebo and drug groups.

## Statistical Criteria

A larger sample size will be needed to detect a significant finding as the alpha level ($p$ value criteria for determining significance) becomes more conservative (i.e., going from maximum acceptable $p$ value of .05 to .01 or .001). This is a potential issue when alpha corrections are needed because of multiple statistical comparisons. A smaller sample size is necessary for one-tailed than two-tailed

statistical comparisons. However, one-tailed statistical comparisons are only appropriate when it is known a priori that any difference between comparison groups is possible in only one direction.

### Sample Size Estimation

There are several commercially available software programs for estimating required sample sizes based on study design, estimated effect size, desired statistical power, and significance thresholds. In addition, free estimation programs might be found through a search using an Internet search engine.

*Ashley Acheson*

***See also*** One-Tailed Test; *p* Value; Two-Tailed Test; Variability, Measure of; Single-Subject Design; Within-Subjects Designs

#### Further Readings

Eng, J. (2003). Sample size estimation: how many individuals should be studied? *Radiology, 227,* 309–313.

Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *The American Statistician, 55,* 187–193.

Polgar, S., & Thomas, S. A. (2008). *Introduction to research in the health sciences* (5th ed.). Toronto, Ontario, Canada: Elsevier.

## Sample Size Planning

Sample size planning is the systematic approach to selecting an optimal number of participants to include in a research study so that some specified goal or set of goals can be satisfied. Sample size planning literally addresses the question "What size sample should be used in this study?" but an answer to the question must be based on the particular goal(s) articulated by the researcher. Because of the variety of research questions that can be asked, and the multiple inferences made in many studies, answering this question is not always straightforward. The appropriate sample size depends on the research questions of interest, the statistical model used, the assumptions specified in the sample size planning procedure, and the goal(s) of the study. In fact, for each null hypothesis significance test performed and/or confidence interval constructed, a sample size can be planned so as to satisfy the goals of the researcher (e.g., to reject the null hypothesis and/or obtain a sufficiently narrow confidence interval). Each of the possible sample size planning procedures can suggest a different sample size, and these sample sizes can be very different from one another.

The most common approach when planning an appropriate sample size is the *power analytic approach,* which has as its goal rejecting a false null hypothesis with some specified probability. Another approach, termed *accuracy in parameter estimation* (AIPE), has as its goal obtaining a sufficiently narrow $(1 - \alpha)\%$ confidence interval for a population parameter of interest, where $1 - \alpha$ is the desired confidence interval coverage, with $\alpha$ being the Type I error rate. Notice that the two perspectives of sample size planning are fundamentally different in their respective goals, with the former being concerned with rejecting a null hypothesis and the latter with obtaining sufficiently narrow confidence intervals. Perhaps not surprisingly, depending on the specific goals, the implied sample size from the two perspectives can be very different. Although other approaches to sample size planning exist, the power analytic and the AIPE approaches serve as broad categories for conceptualizing the goals of sample size planning.

In null hypothesis significance testing, where an attempt is made to test some null hypothesis, not having an adequate sample size can lead to a failure to reject a false null hypothesis—one that should in fact be rejected. When a false null hypothesis is not rejected, a Type II error is committed. In such situations, the failure to find significance oftentimes renders the study inconclusive because it is still uncertain whether the null hypothesis is true or false. Conversely, if the goal of the research study is to reject some specified null hypothesis, then it is not generally a good use of time and resources to use a larger than necessary sample size. Furthermore, if the goal of the research study is to reject some specified null hypothesis and a power analysis suggests a sample size that would be exceedingly difficult to obtain given available resources, a researcher might decide to conduct a modified version of the originally proposed study

or decide that such a study, at present, would not be a good use of resources because of the low probability of success.

A recent emphasis in the methodological literature on effect sizes and confidence intervals has made the dichotomous results of null hypothesis significance testing (reject or fail to reject) less of the focus of the research study than they once were. The AIPE approach to sample size planning attempts to overcome large confidence interval widths by planning an appropriate sample size, where the expected confidence interval width is sufficiently narrow or where there is some desired degree of assurance that the confidence interval width will be no larger than specified (e.g., sample size so that there is 99% assurance that the 95% confidence interval will be sufficiently narrow). Similar in spirit to the AIPE approach is one where an appropriate sample size is planned so that an estimate will be sufficiently close to the population parameter with some specified probability. Still others suggest that power and accuracy should be considered together, where there is some specified probability of simultaneously rejecting some false null hypothesis and obtaining a confidence interval that is sufficiently narrow.

The proverbial question "what size sample should I use" has been repeatedly asked by researchers since inferential statistics became a research tool. However, an appropriate answer depends on several factors and is not always easy to determine. The sample size can be planned for any and all parameters in a model and from multiple approaches. Thus, for research questions that have more than a single parameter, as most research questions do, an appropriate sample size must be based on the question of interest, which is in turn based on one or more parameters. Coupling multiple parameters with different approaches adds to the potential complexity of sample size planning. Regarding multiple parameters, one issue often overlooked is that if several parameters have sufficient statistical power or narrow expected confidence interval width, then there is no guarantee that the entire set of parameters will simultaneously have sufficient statistical power or narrow expected confidence interval width.

An appropriate sample size thus depends on the parameter(s) of interest, the goal of the research study, and the characteristics of the population from which the data are sampled. If there is more than one question of interest, then there will likely be more than one appropriate sample size. In such situations, choosing the largest of the planned sample sizes is advised.

*Ken Kelley*

***See also*** Accuracy in Parameter Estimation; A Priori Monte Carlo Simulation; Confidence Intervals; Effect Size, Measures of; Null Hypothesis; Power Analysis

**Further Readings**

Algina, J., & Olejnik, S. (2000). Determining sample size for accurate estimation of the squared multiple correlation coefficient. *Multivariate Behavioral Research, 35,* 119–136.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Jiroutek, M. R., Muller, K. E., Kupper, L. L., & Stewart, P. W. (2003). A new method for choosing sample size for confidence interval based inferences. *Biometrics, 59,* 580–590.

Kelley, K., & Maxwell, S. E. (2003). Sample size for multiple regression: Obtaining regression coefficients that are accurate, not simply significant. *Psychological Methods, 8,* 305–321.

Kelley, K., & Rausch, J. R. (2006). Sample size planning for the standardized mean difference: Accuracy in parameter estimation via narrow confidence intervals. *Psychological Methods, 11,* 363–385.

Maxwell, S. E. (2000). Sample size and multiple regression. *Psychological Methods, 5,* 434–458.

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology, 59,* 537–563.

## SAMPLING

Sampling occurs when researchers examine a portion or sample of a larger group of potential participants and use the results to make statements that apply to this broader group or population. The extent to which the research findings can be generalized or applied to the larger group or population is an indication of the external validity of the

research design. The process of choosing/selecting a sample is an integral part of designing sound research. An awareness of the principles of sampling design is imperative to the development of research with strong external validity. In theory, a sound sampling method will result in a sample that is free from bias (each individual in the population has an equal chance of being selected) and is reliable (a sample will yield the same or comparable results if the research were repeated).

A sample that is free from bias and reliable is said to be representative of the entire population of interest. A representative sample adequately reflects the properties of interest of the population being examined, thus enabling the researcher to study the sample but draw valid conclusions about the larger population of interest. If the sampling procedures are flawed, then the time and effort put into data collection and analysis can lead to erroneous inferences. A poor sample could lead to meaningless findings based on research that is fundamentally flawed. Researchers use sampling procedures to select units from a population. In social science research, the units being selected are commonly individuals, but they can also be couples, organizations, groups, cities, and so on.

This entry begins by detailing the steps in the sampling process. Next, this entry describes the types of sampling. The entry ends with a discussion of evaluating sampling and determining sample size.

## Steps in the Sampling Process

The sampling process can be diagrammed as shown in Figure 1.

Identifying the population or entire group of interest is an important first step in designing the sampling method. This entire population is often referred to as the *theoretical* or *target population* because it includes all of the participants of theoretical interest to the researcher. These are the individuals about which the researcher is interested in making generalizations. Examples of possible theoretical populations are all high school principals in the United States, all couples over age 80 in the world, and all adults with chronic fatigue syndrome. It is hardly ever possible to study the entire theoretical population, so a portion of this theoretical population that is accessible (the *accessible population* or *sampling frame*) is identified. Researchers define the accessible population/sampling frame based on the participants to which they have access. Examples of accessible populations might be the high school principals in the state of Colorado, couples over age 80 who participate in a community activity targeting seniors, or patients who have visited a particular clinic for the treatment for chronic fatigue syndrome. From this accessible population, the researcher might employ a sampling design to create the *selected sample,* which is the smaller group of individuals selected from the accessible population. These individuals are asked by the researcher to participate in the study. For example, one might sample high school principals by selecting a random sample of 10 school districts within the state of Colorado. In other cases, the accessible population might be small enough that the researcher selects all these individuals as the selected "sample." For example, a researcher studying couples older than age 80 who participate in a particular community activity could choose to study all the couples in that group rather than only some of them. In this case, the accessible population and the selected sample are the same. A third example of an accessible population could be if the patients treated during a certain 3-month time period were chosen as the selected sample from the accessible population of all individuals seeking treatment for chronic fatigue syndrome at a particular clinic.

Finally, the researcher has the *actual sample,* which is composed of the individuals who agree to participate and whose data are actually used in the analysis. For example, if there were 50 older couples at the community activity, perhaps only 30 (a 60% response rate) would send back the questionnaire.

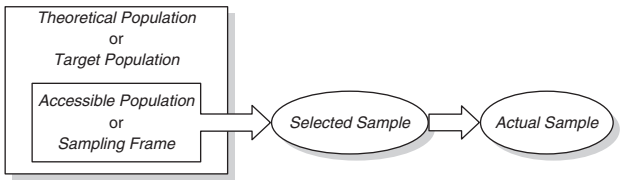



**Figure 1** Sampling Process Diagram

The advantages to using a sample rather than examining a whole population include cost-effectiveness, speed, convenience, and potential for improved quality and reliability of the research. Designing a sound sampling procedure takes time and the cost per individual examined might be higher, but the overall cost of the study is reduced when a sample is selected. Studying a sample rather than the entire population results in less data that need to be collected, which can thereby produce a shorter time lag and better quality data. Finally, by examining a sample as opposed to an entire population, the researcher might be able to examine a greater scope of variables and content than would otherwise be allowed.

Although there are advantages to using sampling, there are some cautions to note. The virtues mentioned previously can also end up being limitations if the sampling procedures do not produce a representative sample. If a researcher is not prudent in selecting and designing a sampling method, then a biased and/or unreliable sample might be produced. Knowledge of what biases might potentially result from the choice of sampling method can be elusive. The errors can be especially large when sample observations falling within a single cell are small. For example, if one is interested in making comparisons among high school principals of various ethnic groups, then the sampling process described previously might prove problematic. A random sample of 10 Colorado school districts might produce few principals who were ethnic minorities because proportionally there are relatively few large urban districts that are more likely to have minority principals. It is possible that none of these districts would be randomly selected.

## Types of Sampling

When each individual in a population is studied, a census rather than a sample is used. In this instance, the accessible population contains the sample individuals as does the selected sample. When sampling methods are employed, researchers have two broad processes to choose from: probability sampling and nonprobability sampling. If using probability sampling, the researcher must have a list of and access to each individual in the accessible population from which the sample is being drawn. Furthermore, each member of the population must have a known, nonzero chance of being selected. Nonprobability sampling, in contrast, is used when the researcher does not have access to the whole accessible population and cannot state the likelihood of an individual being selected for the sample.

### *Probability Sampling*

Probability sampling is more likely to result in a representative sample and to meet the underlying mathematical assumptions of statistical analysis. The use of some kind of random selection process increases the likelihood of obtaining a representative sample. Simple random sampling is the most basic form of probability sampling in which a random number table or random number generator is used to select participants from a list or sampling frame of the accessible population. Stratified random sampling enables the researcher to divide the accessible population on some important characteristic, like geographical region. In this way, each of these stratum, or segments of the population, can be studied independently. Other types of probability sampling include systematic sampling and cluster sampling. The earlier example involving the selection of high school principals used a two-stage cluster sampling procedure, first randomly selecting, for example, 10 school districts then interviewing all the principals in those 10 districts. This procedure would make travel for the observation of principals much more feasible, while still selecting a probability sample.

### *Nonprobability Sampling*

Probability samples, although considered preferable, are not always practical or feasible. In these cases, nonprobability sampling designs are used. As the name would imply, nonprobability sampling designs do not involve random selection; therefore, not everyone has an equal chance of selection. This does not necessarily lead to a sample that is not representative of the population. The level of representativeness, however, is difficult to determine.

A convenience sampling design is employed when the researcher uses individuals who are

readily available. In quota sampling, quotas are set for the number of participants in each particular category to be included in the study. The categories chosen differ depending on what is being studied but, for example, quotas might be set on the number of men and women or employed or unemployed individuals.

## Evaluating Sampling

A variety of steps in the sampling process might lead to an unrepresentative sample. First is the researcher's selection of the accessible population. Most often, the accessible population is chosen because it is the group that is readily available to the researcher. If this accessible population does not mirror the theoretical population in relation to the variables of interest, the representativeness of the resulting sample is compromised. The researcher's choice of sampling design is another entry point for error. It is not very likely that a nonprobability sample is representative of the population, and it is not possible to measure the extent to which it differs from the population of interest. A poor response rate and participant attrition can also lead to a unrepresentative sample. The effects of these nonresponses can be dramatic if there are systematic differences between those individuals who did and did not respond or drop out. Often, there is no way to know to what extent the resulting sample is biased because these individuals might differ from those who responded in important ways.

## Sample Size

There is no single straightforward answer in relation to questions regarding how large a sample should be to be representative of the entire population of interest. Calculating power is the technically correct way to plan ahead of time how many participants are needed to detect a result of a certain effect size. The underlying motivation behind selecting a sampling method is the desire for a sample that is representative of the population of interest. The size of the selected sample depends partly on the extent to which the population varies with regard to the key characteristics being examined. Sometimes, if the group is fairly homogeneous on the characteristics of interest, one can be relatively sure that a small probability sample is representative of the entire population. If people are very similar (as in a selected interest group, individuals with a certain syndrome, etc.), one might suspect the sample size would not need to be as large as would be needed with a diverse group. The level of accuracy desired is another factor in determining the sample size. If the researcher is willing to tolerate a higher level of error for the sake of obtaining the results quickly, then he/she might choose to use a smaller sample that can be more quickly studied. A researcher must also consider the research methodology chosen when determining the sample size. Some methodologies, such as mailed surveys, have lower typical response rates. To obtain an actual sample of sufficient size, a researcher must take the anticipated response rate into consideration. Practical considerations based on time and money are probably the most common deciding factors in determining sample size.

*Andrea E. Fritz and George A. Morgan*

***See also*** Cluster Sampling; Convenience Sampling; Nonprobability Sampling; Population; Probability Sampling; Proportional Sampling; Random Sampling; Sample Size; Stratified Sampling; Systematic Sampling

### Further Readings

Fogelman, K., & Comber, C. (2007). Surveys and sampling. In A. R. J. Briggs & M. Coleman (Eds.), *Research methods in educational leadership and management* (2nd ed.) (pp. 125–141). Thousand Oaks, CA: Sage.

Fowler, F. J. (2002). *Survey research methods*. Thousand Oaks, CA: Sage.

Kraemer, H. K., & Thiemann, S. (1987). *How many subjects? Statistical power analysis in research*. Newbury Park, CA: Sage.

Morgan, G. A., Gliner, J. A., & Harmon, R. J. (2006). Sampling and population external validity. In *Understanding and evaluating research in applied and clinical settings* (pp. 122–128). Mahwah, NJ: Lawrence Erlbaum.

Salant, P., & Dillman, D. H. (1994). *How to conduct your own survey*. New York: Wiley.

Sudman, S. (1976). *Applied sampling*. New York: Academic Press.

# Sampling and Retention of Underrepresented Groups

A fundamental task of research involves the recruitment of individuals to participate in a study. The failure to recruit a diverse sample might affect the size and demographic characteristics of the sample, which might limit the research project in numerous ways. Smaller sample sizes and sample bias might reduce the statistical power of the study and diminish the applicability of findings and interventions based on it. This entry defines underrepresented groups and outlines strategies to enhance the recruitment and retention of underrepresented research participants. All aspects of the research study (i.e., prefieldwork, protocol development, recruitment, pilot research, analysis, and dissemination) are assumed to be interlocking.

## Defining an "Underrepresented Group"

Membership in an underrepresented group might vary depending on the nature of the discipline, the general phenomena being explored, and the particular research study. For example, a review of the psychology literature published by the main journals of American Psychology Association (APA) found U.S. dominance both in authorship and in the sample demographics, thus building an understanding of psychology based on only 5% of the world population and neglecting 95%. Ethnic minorities are overrepresented among studies of the poor and are underrepresented in studies of middle to upper socioeconomic statuses. Historically, studies of risk to the well-being of children and families have not represented the affluent. Studies that focus on a particular racial or ethnic minority group might overly rely on a convenience sample, which might introduce a bias in the sample and perhaps limit the ability to identify the main characteristics and perspectives of the population. An underrepresented group in research, therefore, is defined by the previous lack of representation of the group in research generally or with respect to the particular outcomes and phenomena explored in a given study. Underrepresented groups might include the following: particular racial, ethnic, and socioeconomic groups that might have been represented in the research literature, but perhaps without sufficient representation to reflect the full range of the population (e.g., cultural variation within ethnic groups, low-performing students of high socioeconomic status, or high-performing students of low socioeconomic status); studies that sample exclusively from an underrepresented groups (e.g., emotional-social adjustment of Native American youth with high academic achievement or of Asian-American youth with low achievement); or studies that sample special protected or vulnerable populations such as youth, persons not familiar with English, or persons associated with a socially undesirable characteristic (e.g., homeless, criminal record, or HIV positive). Through a literature review and preliminary fieldwork, researchers should identify the range of the target sample population.

## Strategies to Promote Retention and Engagement of Underrepresented Groups

Strategies to promote the retention and engagement of underrepresented groups include the composition of a diverse research team, prefieldwork and entering into the community, development of measures and piloting the protocol, recruitment and sample maintenance, and steps taken subsequent to data collection. These strategies might be adapted to support various research designs (e.g., qualitative, quantitative, or mixed methods) and might support studies that target underrepresented groups exclusively or that aim to secure a relatively higher percentage or variation of underrepresented groups in a general research project.

### *Composition of Research Team*

Assembling a diverse research team is important to reconcile the assets and liabilities associated with both in-group and out-group affiliations between the research team and the target population.

Although having members of the research team who are themselves members of the underrepresented group might facilitate the recruitment of participants and provide other culturally valued consultations to the research design (e.g., the development of instruments, the interpretation of

results), it is important to note the limitations of in-group members in a research team. Like any research member, in-group members have researcher bias and contribute generally to biases in the data. For example, general concerns about socially desirable responses in the data might be heightened when members of underrepresented groups are being interviewed by a researcher from one's own cultural group. Some respondents might feel more comfortable talking with an outsider because they believe that an outside researcher would be less likely to condemn aspects of the respondents' reported behavior that an insider might consider as failing insider group norms. For these reasons, issues of confidentiality of participants should be reviewed in detail during the informed consent and reiterated in the debriefing statement. Whereas in-group research members might share some key characteristic or perspective, the researcher might be perceived by the underrepresented group as being different in fundamental ways. For example, there might be negative assumptions by the underrepresented group because of the perceived social class of the researcher. Like any other group, underrepresented groups are likely to contain significant variation within the group, so a limited number of in-group researchers would be unlikely to represent the full range of the underrepresented population.

### *Approaches to Entering the Community*

Because participant recruitment typically defines the initial phase of engagement of a participant in a research study, the entry into a community requires culturally sensitive approaches and often a combination of establishing both top-down and bottom-up creditability.

Given that what is relevant to different underrepresented groups might vary both between and across groups, it is crucial for the research team to conduct prefieldwork to become better familiar with the values of a particular target population. Depending on the nature of the study and the desired sample, this prefieldwork might include the following: exploring the extant data (e.g., traditional research, demographic records maintained by government and other sources, or related news coverage), multimedia literature (e.g., books or films produced about and by the target population), and guided and unguided tours of the area. Familiarity with a target community will assist the research team in identifying its official and unofficial leaders.

The preferred approaches to access community leaders vary based on the nature of the research study, the target population and related institutions, and the organizational affiliations of the research team. A research study that focuses exclusively on a particular cultural group in a defined geographic region might use approaches that are typical to ethnographic research, such as frequent attendance at community events and formal meetings with key officials either as individuals or as invited members of an advisory group. In such scenarios, approval by one community leader might help garner additional support for the study from other leaders and eventually from potential participants. Research designs that seek to increase the participation rates of underrepresented groups in more general studies need to establish such credibility with multiple groups or identify institutions (e.g., schools, neighborhood programs that serve diverse youth and families, or laundromats) that serve as gateways for multiple underrepresented groups.

To forge formal relationships with a target research population, the research team must articulate a comprehensive overview of the research project in a concise and compelling manner. The research team should be prepared to provide brief written materials about the research project to the decision makers, as well as the materials that would be distributed to potential participants or their parents or guardians. In addition to using culturally sensitive language that might include translated materials, the materials should clearly indicate the affiliations of the organizations sponsoring the research (e.g., university, government, service agency, or nonprofit) as well as any funding sources. Certain affiliations might provide legitimacy in gaining official approval to enter a community or organization. Alternatively, certain research affiliations might serve as barriers to the target population because of suspicion about the motives of the research, particularly given the funding sources. Affiliations of the research project and their related outcomes (i.e., whether findings would be published that conflict with the mission of a collaborating organization or report illegal

behavior) might complicate the process of accessing an underrepresented or vulnerable group.

Often in coordination with gaining formal entry into a group, a research team might be engaged in developing informal entry points. Informal community leaders and informants (e.g., community elders, school secretaries, or barbers or beauticians) often provide instrumental advice regarding the development of measures related to the phenomena being studied, as well as strategies to secure additional approval and participation in the study.

### *Development of Measures and the Protocol*

Insights gathered through prefieldwork, consultation with formal and informal leaders, and pilot research in a community might substantively inform aspects of the research protocol including measures, methods, and incentives.

Prefieldwork and communication with members of the target community might assist in the development of a more culturally sensitive protocol. For example, such interaction might lead to changes in word choice in a questionnaire to make it better understood by the target population. Members of the target population might suggest questions or follow-up prompts to explore nuances of a subject that the research team might have been unaware of but are salient to the target population. In addition, pilot research might suggest that significant changes need to be made in the delivery of the protocol. For example, a traditional paper/pencil questionnaire might be read to participants or administered via an audio-enhanced computer program to decrease the cognitive and language load for participants. The decrease in the cognitive and language demands might increase the overall validity of responses, as well as increase participation rates of individuals who might have limited English proficiency, have lower reading levels, or are generally younger. Such adjustments to the administration of the research protocol might decrease the amount of time required to participate, thus encouraging the involvement of more participants.

Incentives might be both intrinsic and extrinsic. Depending on the nature of the study (e.g., studies that explore positive development), members of underrepresented groups might feel validated simply to contribute their perspective to research and to research-based applications and policies. The nature of the study and its research design (e.g., longitudinal or cross-sectional) have implications on the frequency and amount of incentives. Informal community members might help identify what type of incentive would be both desirable for participants and within the guidelines established by the community and Institutional Review Board. In terms of material incentives, research projects might consider provisions to the institutions that facilitate the research study at their organization and also individual incentives for the participants. Even if the study is cross-sectional, it is important to maintain positive relations with the facilitating organizations so that they might be willing to assist other research projects in the future. For longitudinal studies in which the success of the research is predicated on being able to maintain the same participants throughout the study, there might be incremental incentives where the participants receive more of the incentive for each wave of data collection, perhaps with a "balloon payment" for the final data collection or entry into a raffle.

### *Recruitment and Sample Maintenance*

When compared with less diverse samples, the time frame and budget for a research project that seeks to increase the number of participants from underrepresented groups might require significantly more resources. Researchers might pursue recruitment possibilities that might not yield a high number of participants but might enrich the data because the sample might move beyond convenience and snowball to include purposeful sampling. To address prospective participants' feelings of suspicion and doubt that might exist regarding research, the relationship building required to earn top-down and bottom-up credibility requires time. Although some institutions might have the resources to assist in recruiting participants for a study with minimal researcher involvement, other institutions might be consumed with their own daily operations. For the latter, to secure participants, the research team will likely need to make numerous site visits first to negotiate institutional approval and then to secure consent forms and conduct the research. If minors are involved,

this process might require additional negotiation and interfacing by phone or in person with parents/guardians. The research team needs to convincingly communicate (a) why the research project is important generally, (b) why their participation in the project is valuable specifically, (c) what risks there are in participation, (d) the personal and collective benefits of participating for the individual and the institution, and (e) the benefit of contributing to scientific knowledge and the cost of not having their perspective represented in research-based funding and policy decisions.

### *Steps Taken Subsequent to Data Collection*

The parts of the research project are iterative, as should be the communication process with participants and the research and practice community. The role of the participants in the data analysis might depend on the nature of the research. In participatory action and collaborative action research, participants reflect on the data and its interpretation, although participants might play no direct role in data analysis in other research designs. In all research, participants should be part of the dissemination loop; how this is determined might vary and include a research summary written in nontechnical language, such as newsletters for participants, site reports, and other consultations that might be negotiated based on the nature of the project. For example, some organizations might be interested in including data collected by the research team in their own assessment of their program and services.

*Mona M. Abo-Zena*

***See also*** Qualitative Research; Quantitative Research; Recruitment; Sampling

### Further Readings

Arnett, J. J. (2008). The neglected 95%: Why American psychology needs to become less American. *American Psychologist, 63,* 602–614.

Cauce, A. M., Ryan, K. K., & Grove, K. (1998). Children and adolescents of color, where are you? Participation, selection, recruitment, and retention in developmental research. In V. McLoyd & R. Sternberg (Eds.), *Studying minority adolescents.* Mahwah, NJ: Lawrence Erlbaum.

Luthar, S. S., & Latendresse, S. J. (2005). Children of the affluent: challenges to well-being. *Current Directions in Psychological Science, 14,* 49–53.

Patton, M. Q. (1980). *Qualitative evaluation methods.* Beverly Hills, CA: Sage.

Weiss, R. S. (1994). *Learning from strangers: The art and method of qualitative interview studies.* New York: The Free Press.

## Sampling Distributions

Sampling distributions are the basis for making statistical inferences about a population from a sample. A sampling distribution is a set of samples from which some statistic is calculated. The distribution formed from the statistic computed from each sample is the sampling distribution. If the statistic computed is the mean, for example, then the distribution of means from each sample form the sampling distribution of the mean. One problem solved by sampling distributions is to provide a logical basis for using samples to make inferences about populations. Sampling distributions also provide a measure of variability among a set of sample means. This measure of variability will, in turn, allow one to estimate the likelihood of observing a particular sample mean collected in an experiment to test a hypothesis.

At the simplest level, when testing a hypothesis, one is testing whether an obtained sample comes from a known population. If the sample value is likely for the known population, then it is likely that the value must come from the known population. If the sample value is unlikely for the known population, then it likely does not come from the known population, and it can then be inferred that it comes from a different unknown population instead. If some treatment is performed, such as giving a drug to improve patient recovery time, then the average recovery time from a sample of the treated group will allow a test of the idea that the treatment had some effect (here on recovery time). Does giving patients this new drug in effect create a new and different population—a population using the drug? If the average recovery time of the treated group is very similar to, or likely for, the known population of patients that do not take the drug, then the treatment likely had no effect. If

the average recovery rate is very different from, or very unlikely for, the known population of patients not taking the drug, then the treatment must have had an effect and created a new population of patients with different recovery times. Thus, some way to judge how likely a value is for the known population is needed.

The common formula used to find the probability or the likelihood of a value for a known population (solving *z*-score problems) is

$$z = \frac{X - \mu}{\sigma}.$$

In the previous formula, the standard deviation sigma ($\sigma$) provides information about how much variability exists in the population. Knowing how much variability exists in the population (the width of the distribution of scores) allows one to know how likely a single $x$ value is for that population. Because most values in a distribution will lie close to the mean, less likely values will fall farther from the mean. The wider the distribution of scores, the less pronounced any specific difference between a value and the mean will be. For example, if the difference between an $x$ value and the population mean remains constant (in the numerator), then that difference will be much more likely if the population has a very wide distribution (large denominator) compared with its likelihood in a very narrow distribution (small denominator). So, any factor, like a decreased spread in the distribution of scores, which increases the relative difference between a value and the mean will lower the estimate of how likely the value is for the distribution.

However, when testing a hypothesis, it is never based on a single $x$ value. Instead, a sample of values is used from which the average is computed. If the average or mean value tested is very different from the known population, then it can be assumed that the population the sample represents is not the same as the known population. The problem in using the previous formula is that sigma gives information about how much individual values vary within a population but provides nothing about how much *sample means* vary. Sampling distributions provide an explanation of how to measure variability in samples and, thus, the probability of observing a particular sample mean.

## Sampling Distribution of the Mean

Sampling distributions are theoretical and not actually computed. However, examining the process of computing one is necessary. There are many types of sampling distributions, and a sampling distribution for any statistic can be formed. For the current discussion, the sampling distribution of the mean is most relevant. To form a sampling distribution, a researcher will do the following:

1. Sample repeatedly and exhaustively from the population.
2. Calculate the statistic of interest (the mean) for each sample.
3. Form a distribution of the set of means obtained from the samples.

The values taken from the population to form a sample can be any specific size, but every possible sample of that size from the population must be taken. Then, an average of each sample is computed. The set of means obtained from each sample will form a new distribution, which is called a sampling distribution. In this case, where the mean is computed as the statistic, it will be the sampling distribution of the mean. Every possible combination of values from the population is sampled to form a true sampling distribution. Because most populations are very large, it is impractical actually to go through the process of forming a sampling distribution, which is why they remain theoretical.

## Information Gained

The first important fact learned from the sampling distribution of the mean is that the mean of the population and the mean of the sampling distribution of means will have exactly the same value. That is, the average of the entire population of single $x$ values is exactly the same as the average value obtained if each sample mean from the sampling distribution of means is averaged together. This fact is important to hypothesis testing because when testing a hypothesis based on a sample, even though a single sample will not likely yield a mean exactly the same as that of the population, on average it will be

exactly the same. Thus, it is certain that repeated experiments will yield samples that will on average be the same as the population. Using a sample to make an inference about a population is therefore a logical and reasonable proposition.

The next important piece of information obtained from the sampling distribution of the mean is a measure of variability among sample means. Recall that some way to measure how much variability exists in a set of sample means is needed so that there will be some way to gauge how likely it is to obtain a particular sample mean collected in an experiment. If the mean value obtained in a sample is unlikely for the known population, then the population the value comes from is probably different from the known population. If the mean value obtained in the sample is likely or similar to the known population, then it is likely there is no difference between the population the value comes from and the known population. If there is a large amount of variability from sample to sample, then an individual sample mean obtained to test a hypothesis will have to be much more different from the known population to stand out distinctively and be considered unlikely for that population than if there is little variability from sample to sample. The standard deviation, just as with other distributions, will be the measure used to indicate the spread or dispersion of a distribution. Just as with the standard deviation of a population of individual values in which the amount of variability that exists is measured by how much individual values deviate from the average, variability in the set of sample means will be computed the same way. When measuring the average deviation of a set of sample means in a sampling distribution, the amount of variability that exists from sample to sample is being measured. This measure of the standard deviation of a distribution of sample means is called the *standard error* and is symbolized as $\sigma_{\overline{X}}$.

Because the sampling distribution of the mean is theoretical, there is no need to calculate the standard error every time an inferential test is conducted. Instead, an estimate is made from the population or the sample. The formula to estimate the standard error from the population is

$$\sigma_{\overline{X}} = \sigma/\sqrt{n}.$$

The previous formula can be used if the population standard deviation is known. If so, it forms the denominator for a $z$-score hypothesis test:

$$z = \frac{\overline{X} - \mu}{\sigma_{\overline{X}}}.$$

If the population standard deviation is not known, which is usually the case, then the population standard deviation must be estimated from the obtained sample. Although the standard deviation estimated from a sample is calculated slightly differently than when all the population values are known, the computation of the standard error is essentially the same. In such cases, the standard error is usually represented with Roman instead of Greek letters. So, the standard error is represented as $S_{\overline{X}}$ and is computed with the formula

$$S_{\overline{X}} = S/\sqrt{n}.$$

In addition, when using the sample to estimate the standard deviation, one is no longer computing a $z$ test but a $t$ test instead:

$$t = \frac{\overline{X} - \mu}{S_{\overline{X}}}.$$

Notice that the denominator is an estimate of the standard error, and it is the same whether computing a $z$ test or a $t$ test. The distance a sample mean falls from the mean of the population is mediated by how much variability exists from sample to sample. If it is relatively unlikely to observe a certain sample ($p < .05$ for $\alpha = .05$), then one can conclude that the sample did not come from the known population.

Finally, sampling distributions also yield information about how large a sample needs to be to test a hypothesis. The shape of the sampling distribution of the mean will always be normal *regardless* of the shape of the population distribution. Whether the population distribution has a normal, positively or negatively skewed, or unimodal or bimodal in shape, the sampling distribution of the mean will always have a "normal" (unimodal and symmetric)

shape. That is because when a distribution of sample means is formed, each value in the distribution is derived from a sample that contains a variety of scores from the population. Because each value in the sampling distribution is an average of these values from the population, most of the scores will be close to the mean of the population and create unimodal and symmetric distribution, even if the values in the population of single $x$ values do not.

Recall that when values are used to form a sampling distribution, samples of any size can be used. However, the larger the number of values in a sample taken from the population to form the sampling distribution, the more "normal" the sampling distribution will be. That is because there will be a larger variety of values from the population in any individual sample, and it is more than likely the average from each sample will approximate the average value of the entire population. Thus, when 30 values in a sample are collected, enough variety is contained in the sample for those values to average out very close to the average of the population. However, the larger the number of values one takes in a sample, the closer one gets to the average of the population. Because an estimate of the standard error is usually made from a sample, the sample size needs to be around 30 to approximate the value that would be obtained if the standard error was computed from the population. Thus, the minimum number of values needed to approximate the population with a sample is close to 30, and it is best to have this minimum number in any sample used for hypothesis testing.

*David S. Wallace*

***See also*** Central Limit Theorem; Normal Distribution; Standard Error of the Mean; *t* Test, One Sample; Variance

**Further Readings**

Gravetter, F. J., & Wallnau, L. B. (2002). *Essentials of statistics for the behavioral sciences* (4th ed.). Pacific Grove, CA: Wadsworth.

Hays, W. (1994). *Statistics* (5th ed.). Orlando, FL: Harcourt Brace.

Howell, D. C. (1999). *Fundamental statistics for the behavioral sciences* (4th ed.). Pacific Grove, CA: Duxbury Press.

# Sampling Error

Sampling error addresses how much, on average, the sample estimates of a study characteristic or variable, such as years of education, differ from sample to sample. Sampling error is essential in describing research results, how much they vary, and the statistical level of confidence that can be placed in them. Sampling error is also critical in tests of classic statistical significance. This entry provides basic definitions of concepts inextricably related to sampling error, describes when it is appropriate to calculate sampling error, and outlines when complications might occur in these calculations.

## Populations and Samples

Which candidate do voters prefer in an upcoming election? What is the average level of formal education among American adults at least 25 years old? Considerable research tries to ascertain the value of some characteristic in a well-defined population. Researchers call the value of a selected characteristic in the population, such as a preference for "Candidate X," a *population parameter.*

A *population* is the entire collection of cases, elements, or people that the researcher plans to study, such as all U.S. registered voters in November 2008. Notice the precision in the example population definition—it is specific to a particular quality such as voter status, location, and time. If a population is small, cooperative, and accessible, then collecting information is straightforward. For example, a small county has an address list of all registered voters. Such a list is called a population frame, and researchers ensure that it is as complete as possible. Collecting information from all population elements is called a census.

Most often, however, the population is too large, geographically dispersed, or unwieldy to collect information efficiently from every single element. When the United States conducts its decennial population census, it can take a year to collect the information; even with such a massive effort, the U.S. Census undercounts some population segments, spending additional time making corrections. Thus, much research gathers

information from a *sample* (i.e., a specified subpart or subset of the population). The value of a selected characteristic in a sample is called the *sample estimate* or the *sample statistic*. For example, surveys are gathered about voter preferences using samples prior to elections. Sampling error forms an integral part of generalizing from a sample to the larger population because the exact population value is typically unknown.

Many studies predict the population parameter from a sample estimate. However, because sample estimates rely on only a subset of population cases, the estimates will vary across the samples. If the sample is carefully selected using probability methods comparable with a lottery, most estimates will be close to the population parameter, but a few will deviate sharply from the population value. The study of this sample variability comprises much of the study of sampling error.

## Types of Error in Parameter Estimation

Research practitioners distinguish two major sources of error in estimating population parameters. First is systematic error or *bias,* which is often hidden and thus uncontrolled. An example is the scale that always weighs its user 5 lb too light. Poundage loss or gain is reflected in the scale results—but always 5 lb under the true weight. Given a strong focus on sampling error, bias is often critically neglected. Strategies used to minimize bias include careful measurement (e.g., a correctly calibrated scale), using several different measures, and a probability sample.

The second major source of error in estimation is sampling error, or the average estimated variability of results from sample to sample. It is random and can sometimes be estimated in advance of the study if the researcher has sufficient information about the population. Random error means exactly that: There are no systematic deviations from the population parameter across a potential range of samples. Fluctuations in one direction from a sample mean or average in the long run over many samples are assumed to cancel fluctuations in a different direction from other samples. The results from any single sample, of course, could deviate in *any* direction from the population value. By definition, sampling error should only be calculated on probability samples because the laws of probability are used in its estimates.

## Basic Types of Samples

Probability samples, in which each element or case has a known, nonzero chance of selection from a well-defined population, differ from nonprobability samples, in which the chances of selection are unknown. Typically, nonprobability samples have unknown chances of selection because human judgment has been substituted for probability selection methods. For example, students gathering data for a project might survey all customers visiting a convenient grocery store. Many magazines gather information from self-selected readers who mail in a completed questionnaire. Neither sample represents its larger population of grocery customers or magazine readers.

## Effects of Sample Type on Sampling Error

Novice researchers are most familiar with simple random samples (SRS); not only does each element have an equal chance of selection, but also each combination of elements is equally probable. Think of a state lottery in which each winning number in a sequence is mechanically and separately drawn making any number progression possible. SRS calculations are the basic formulas used for sampling error in most statistical computer programs, such as the Statistical Package for the Social Sciences (SPSS, an IBM product). Calculation formulas for sampling error on other equal probability samples, such as cluster samples, in which several adjacent cases are selected at once (as, for example, when a researcher collects data from an entire classroom), are more complex.

Not all probability samples employ equal selection of sample elements. Some employ unequal selection: For example, to obtain sufficient women physicists in a study of university faculty, a researcher might "oversample" females, giving them greater chances of entering the study than men. Again, formulas to calculate sampling errors in unequal probability samples are more complicated.

Samples drawn from large, dispersed populations, especially where an initial complete frame is unavailable, often use multistage designs. In an

area sample of households, researchers might first select states, then counties or parishes, blocks within counties, dwelling units within blocks, and respondents within dwelling units. Calculating sampling errors for multiple-stage designs is more complicated than those for a single-stage design.

## Sampling Error and Sampling Distributions

An important statistical construct behind estimating sampling error is the *sampling distribution,* which is a (usually hypothetical) set of sample estimates from a theoretically infinite number of samples of the same size and type gathered within about the same time period. When pollsters draw many samples from a pool of registered voters during an election season, the data actually begin to approximate a sampling distribution: Suppose many samples were drawn from registered voters using SRS of size 1,600, with each respondent indicating his or her candidate choice. Suppose 40% preferred Candidate X in each of five samples, 30% chose her in one sample, 45% chose her in each of three samples, and so on. We then could create a graph with Candidate X preference percentage along the horizontal ($x$) axis and the frequency with which each sample estimate occurred along the vertical ($y$) axis. Given many such samples and many respondents in each one of those samples, the distribution of sample percentage estimates should resemble the familiar bell-shaped or normal curve. The average or mean percentage of all the sample results (40% in this example) is considered an *unbiased estimator* of the population parameter—what the researchers would have obtained had they been able to contact all U.S. registered voters within a short time period. It is critical to remember in estimating sampling error that the unit in a sampling distribution is a single sample result rather than an opinion from an individual respondent.

## Confidence Intervals and Standard Errors

Considerable research predicts the population value from a sample statistic. It is helpful to approximate how much estimates can be expected to vary across samples. If the characteristic or variable of interest is numeric (e.g., years of age or dollars of income), then positive and negative error limits, which are called *confidence intervals,* can be placed around a single sample statistical estimate. Confidence intervals can only be constructed if the characteristic is an actual number or percentage. Most surveys or polls report confidence intervals when they explain that study results can vary by a certain amount (e.g., "40% of those interviewed prefer Candidate X—plus or minus 5%." Thus, support for Candidate X could range from 35% to 45%—in most samples—typically 95 of 100 probability samples. Four main components comprise the confidence interval, as follows:

1. The result from one sample (e.g., 40% choosing Candidate X for President.
2. The variability around the mean, percentage, or average within that single sample called the *standard deviation.*
3. The size of the sample; recall that each sample size in the sampling distribution is the same (e.g., 1,600).
4. How confident the researcher wishes to be in his or her estimate. Sample estimates vary; some are very close to the population parameter (say, 39%), and a few are very different (e.g., a sample in which only 16% support Candidate X). The typical choice is to construct the interval such that 95% of the samples drawn produce estimates relatively close to the population parameter.

The estimate of average variability around the population parameter is called the *standard error* or sometimes the *standard error of the estimate.* The standard error is a measure of variability for the sampling distribution (i.e., how much we expect the results to vary on the average from sample to sample and from the "average sample" to the population parameter). If the sampling distribution approximates a normal curve, which it usually does with samples that have large casebases, by definition 95% of the sample estimates will be within two standard errors on either side of the estimated population parameter. In contrast, 5% of the samples will deviate considerably from the actual population parameter. Such an erroneous estimate can occur even with a probability sample. Based on the laws of chance, every now and then, there is a totally unrepresentative sample.

The results from large samples vary less from one sample to another than results from small samples: Larger samples have smaller sampling errors, thus producing more stable predictions. In fact, one way to minimize (thus partially control) sampling error is to take much larger samples. Samples with very little internal variability (small standard deviations) also differ less from sample to sample. So either large samples or samples with low internal variability produce estimates that deviate less across each another, have smaller confidence intervals, and tend to approximate the estimated population parameter more closely.

Putting it all together, the generic calculation formula for the 95% confidence interval (CI) around the mean is

$$\text{CI} = \overline{X} + 1.96 \times \sigma_{\overline{X}},$$

where $\overline{X}$ is a sample mean, $+1.96$ is the number of estimated standard error of the mean units on either side of mean ($-1.96$ to $+1.96$ standard errors incorporate about 95% of sample results), and $\sigma_{\overline{X}}$ is the estimated standard error of the mean.

Although it is somewhat cumbersome to do, the confidence interval can be estimated with a calculator, and virtually all statistical computer programs calculate it in seconds.

### Some Other Uses of Sampling Error

Sampling error is important in creating estimates of the population value of a particular variable, how much these estimates can be expected to vary across samples, and the level of confidence that can be placed in the results. Furthermore, estimates of sampling error are involved in tests of statistical significance for most basic classic statistics. Some form of sampling error serves as a denominator in *t* tests for the difference between means from two separate samples, in analysis of variance and in regression statistics. Sampling error reminds us that it is not sufficient simply to "eyeball" differences in means or percentages across different groups because these differences might simply reflect chance sampling fluctuations rather than "real" differences. The study of sampling error reaffirms that sample statistics are just that: estimates of the population value from particular collections of data.

*Susan Carol Losh*

***See also*** Confidence Intervals; Nonprobability Sampling; Normal Distribution; Probability Sampling; Sampling; Sampling Distributions; Standard Error of Estimate

#### Further Readings

Babbie, E. (2004). *The basics of social research* (10th ed.). Belmont, CA: Thomson Wadsworth.

Kalton, G. (1983). *Introduction to survey sampling: Quantitative applications in the social sciences #35*. Thousand Oaks, CA: Sage.

Kish, L. (1965). *Survey sampling*. New York: Wiley.

## SAS

SAS, which was originally called the Statistical Analysis System, is a statistical software package that runs on both PC and UNIX platforms as well as on other mainframe computers. Growing out of a project aimed at management and analysis of agricultural experiments in the Department of Experimental Statistics at North Carolina State University in the late 1960s, SAS has evolved into an integrated system of data management and decision support tools. SAS can perform data entry, retrieval, management, and mining; graphics generation and reporting; statistical analysis; business planning, forecasting, and decision making; operation research and project management; and quality improvement. Although SAS offers a point-and-click front end (the windows-based graphic user interface [GUI]) to users not familiar with the SAS language, SAS programming provides more flexibility, efficiency, and options in statistical analysis. The first limited release of the SAS system was SAS 71 and the latest version, as of January 2009, is SAS 9.2.

### Components

SAS consists of several components that might require separate licenses and installation. A list, although it is far from being complete, includes the following: Base SAS (the core of SAS),

Enterprise Miner (a data mining tool), SAS/ACCESS (for the transparent sharing of data with non-native data sources), SAS/AF (application facility), SAS/CONNECT (for communication between different platforms), SAS/ETS (for econometric and time series analysis), SAS/GRAPH (for charting on graphical media), SAS/INSIGHT (a dynamic data mining tool), SAS/IML (an interactive matrix language), SAS/OR (for operations research), SAS/QC (quality control tools), SAS/SHARE (a data server which enables multiple users to have simultaneous access to SAS files), and SAS/STAT (for statistical analysis).

## Features

SAS has many features that distinguish it from other statistical software. It can read, write, and process data of different formats. Its system of formats and informats, which are used to control representation and categorization of data, enables users to create customized user formats. SAS also provides interaction with the operating system. For example, in Microsoft Windows, through the Dynamic Data Exchange (DDE), users can connect SAS results with Microsoft Excel.

It also supports the implementation of Structured Query Language (SQL), which is a standardized, widely used language that retrieves and manipulates data in relational tables and databases. SQL provides an alternative to other SAS procedures or the DATA step and is an easy and flexible way to query and combine data.

It supports dynamic data-driven code generation through the SAS Macro language, which can make the SAS programs more dynamic and reusable. Large programs can be simplified by creating functions with parameters that can be evoked several times. Macros and other SAS commands can be stored in an external file that can be shared and used easily by other users. The macros, to some extent, are similar to the packages or toolboxes in other statistical software and minimize the tedious work on programming and significantly facilitate statistical planning, research, and preparation of results.

The Output Delivery System (ODS), which is designed to overcome the limitations of traditional SAS output, can deliver the output in various formats, such as RTF, PDF, HTML, XLS, XML, and SAS data sets. ODS provides high-quality, professional-looking and detailed presentation output from SAS. Various styles available in ODS also enhance the presentation output by controlling the report's content, overall color scheme, font, and so forth.

Recent releases of SAS version 9 introduced new ways to create statistical graphics using the ODS Statistical Graphics (or ODS Graphics for short), which enhances its graphics that have been criticized when compared with other statistical software. ODS Graphics eliminates the need to save numerical results in an output data set, manipulate them with a DATA step program, and display them with a graphics procedure. ODS equips many SAS procedures to create graphics as automatically as tables without additional programming.

SAS/ACCESS modules allow communication with databases. Database tables, in most cases, can be treated as if they were native SAS data sets. As a result, data from different platforms can be combined without knowledge of the details and distinctions between data sources.

The Interactive Matrix Language (SAS/IML) gives users access to powerful and flexible programming in a dynamic, interactive environment. The results can be observed immediately; alternatively, statements can be stored in a module and executed later. SAS/IML makes possible easy and efficient programming with the many features for manipulation of arithmetic and characteristic expressions. As a part of the SAS system, it also gives users access to SAS data sets or external files with the rich SAS commands for data management.

## Application

Partly because of its powerful capability to import, export, and manipulate data sets of various formats, SAS is widely used for statistical analysis in many fields, such as the insurance industry, public health, scientific research, finance, human resource management, IT service management, utilities industries, retail industry, and so forth. SAS continues to be the standard statistical analysis software used for the submission of reports of clinical pharmaceutical trials to the U.S. Food and Drug Administration.

### SAS User Groups

SAS has various local, regional, international, or professional user groups that promote communication about understanding and using the software more productively. These SAS user groups make available some of the resources and the expertise of SAS programmers to the new users of SAS. SAS Users Group International (SUGI) is a nonprofit organization open to all SAS software users throughout the world. SUGI plans and sponsors an international annual conference to exchange ideas and explore ways of using SAS.

There are other regional users groups, such as the six regional users groups in the United States as well as users groups that focus on special interests, such as the Business Intelligence SAS Users Groups (BISUG). The online SAS users group, SAS-L, is a worldwide online community for SAS software users. SAS-L is essentially a listserv e-mail list, which can be obtained either through Usenet, Web interface, or subscription to the e-mail list.

### SAS in Research Design

The construction of optimal research designs requires powerful algorithms. SAS provides good support of research design, including procedures such as FACTEX, OPTEX, and PLAN, which provide researchers with tools for the construction and optimization of research designs. The ADX interface (available in SAS/QC) provides an easy-to-use and intuitive interface, easy data entry, greater facility, and interactive graphics throughout the entire research design process. In addition, a collection of macros exists that can be obtained as a programming interface to tackle common research design tasks.

Sample size requirements can be obtained easily using the POWER or GLMPOWER procedures. The FACTEX procedure constructs orthogonal factorial experimental designs, and the OPTEX procedure searches for optimal designs in situations in which standard designs are not available. The PLAN procedure can be used to construct designs and randomization plans for factorial experiments, especially nested and crossed experiments and randomized block designs. The new release of SAS version 9.2 includes an experimental procedure SEQDESIGN, which provides various methods to compute the boundary values for a group sequential design. These procedures, together with the ADX interface, provide a rich set of design tools for both proficient SAS programmers and those who are new to SAS programming.

### Disadvantages

Although SAS has many features and is powerful for both data manipulation and analysis, it has some disadvantages when compared with other statistical software. Licensing must be purchased, although free alternatives exist, such as R. It requires programming, which takes time to learn, to explore fully the many options it provides. It has been criticized for having graphical output inferior to other statistical software, such as Mathematica.

*Jingyun Yang and Xuebin Wang*

***See also*** Factorial Design; Sample Size; Sequential Design; SPSS; SYSTAT

#### Further Readings

Atkinson, A. C., Donev, A. N., & Tobias, R. D. (2007). *Optimum experimental designs, with SAS*. New York: Oxford University Press.

Delwiche, L. D., & Slaughter, S. J. (2003). *The little SAS book: A primer* (3rd ed.). Cary, NC: SAS Institute.

Rodriguez, R. N. (2004). An introduction to ODS for statistical graphics in SAS 9.1. Presented at the SUGI 29 proceedings, Montréal, Canada.

## SCATTERPLOT

A scatterplot is a graphic representation of the relationship between two or three variables. Each data point is represented by a point in $n$ space, where $n$ is the number of variables. The most common type of scatterplot involves two variables, with data indicated by its bivariate coordinates, usually denoted by $X$ and $Y$. Trivariate plots are also common. Representing data in more than three dimensions requires multiple bivariate and/or trivariate plots. So basically, all scatterplots fall into two types: bivariate and trivariate.

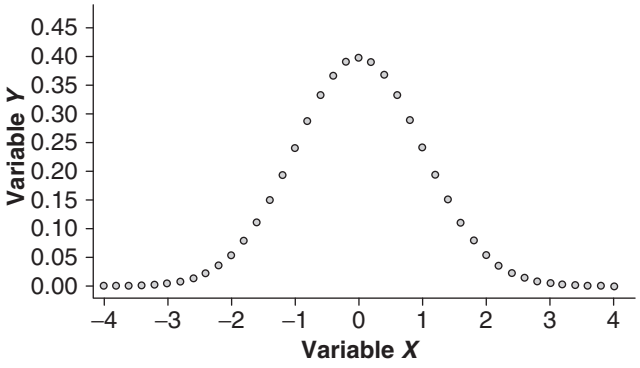



**Figure 1** Bivariate Scatterplot

Manipulated data points are used to present a bivariate scatterplot in Figure 1 and a trivariate plot in Figure 2.

## Scatterplots in Data Analysis

Scatterplots can be used to represent variables that have linear, nonlinear, or no relationship. Often scatterplots can be used to provide researchers with the information necessary to decide whether they should fit a linear model to their data. This is particularly important if they plan to use a statistical technique that assumes linearity, such as an ordinary least squares regression. Looking at the scatterplot in Figure 3 would lead a researcher to understand that fitting a linear model would be misleading.

### *Correlation and Regression*

Correlation is a good way to examine the linear relationships between two variables, for example, a person's weight and height, or a student's high school grade point average, and his SAT/ACT score. The strength of the relationship between two variables is usually described in terms of the correlation coefficient (also known as the Pearson correlation coefficient), which ranges from −1 to 1. A scatterplot is often used to provide researchers the graphic view of what tends to happen to one score when another score increases/decreases.

When a set of variables is at hand, it is fairly easy to draw the scatterplot by plotting one score on the vertical axis and the other on the horizontal axis. Figures 4 through 7 provide examples of variable pairs with correlations of −0.8, −0.3, 0.3, and 0.8, respectively.

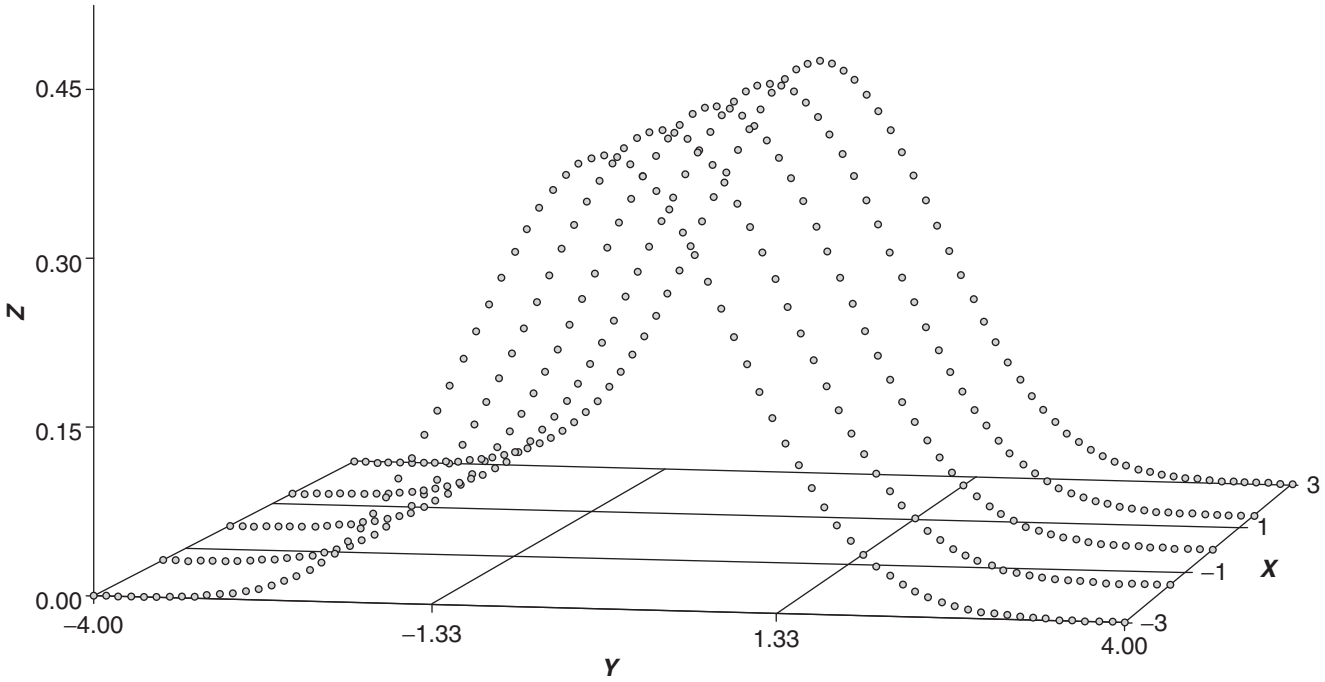



**Figure 2** Trivariate Scatterplot

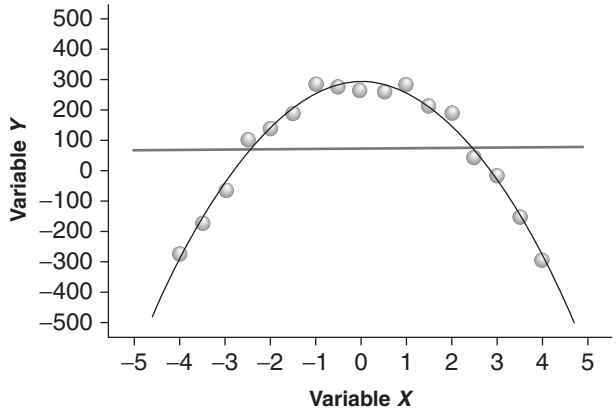



**Figure 3** Nonlinear Model With Zero Correlation

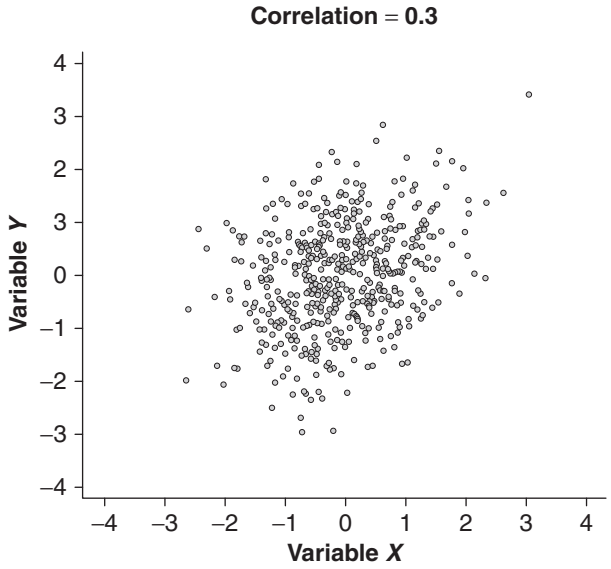



**Figure 4** Bivariate Linear Relation With Positive Correlation = 0.3

A positive correlation describes the situation in which an increase in variable $X$ is associated with an increase in variable $Y$, whereas a negative correlation implies that an increase in variable $X$ is associated with a decrease in variable $Y$. But a correlation of 0.8 is not stronger than a correlation of −0.8. It is the magnitude that matters. They simply work in opposite directions.

In certain extreme situations, all of the dots fall on a straight line. This is called a perfect correlation. Figures 8 and 9 show perfect positive and perfect negative correlations, respectively.

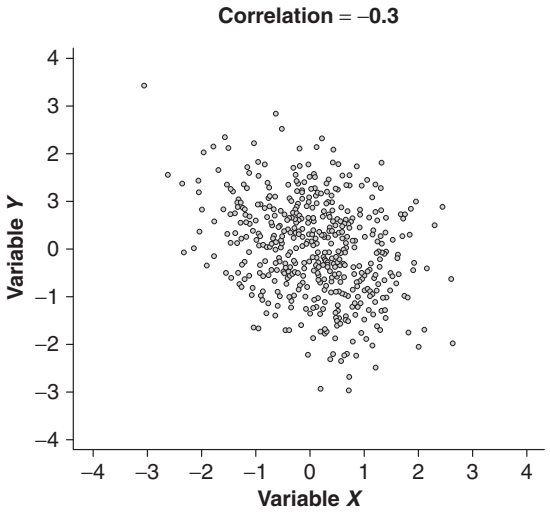



**Figure 5** Bivariate Linear Relation With Negative Correlation = −0.3

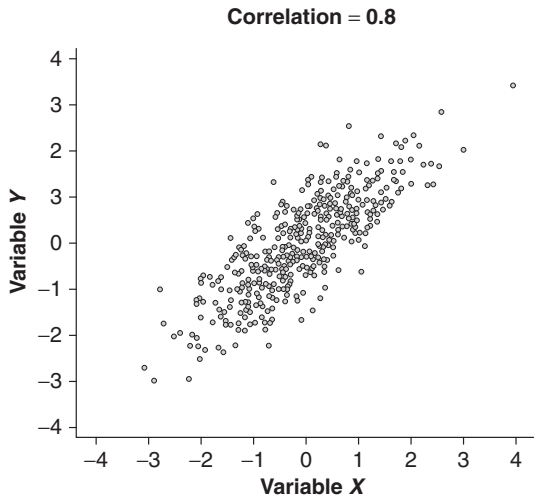



**Figure 6** Bivariate Linear Relation With Positive Correlation = 0.8

The line interpolating all the dots in either Figures 8 and 9 is considered as the line of best fit. In a real-world data analysis, the line of best fit does not necessarily go through all the dots in a scatterplot. Essentially, the line of best fit means that a line is closest to most of the dots or a line is as close to most of the dots as possible. The vertical distances between the line and those dots are called residuals. In statistics, the least squares

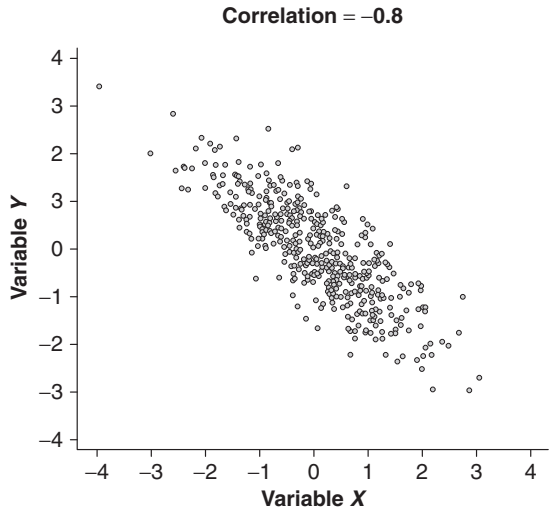



**Figure 7** Bivariate Linear Relation With Negative Correlation = −0.8

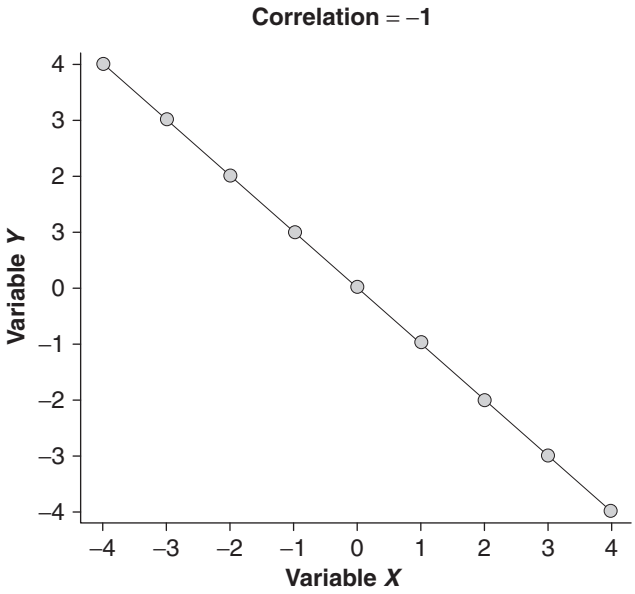



**Figure 9** Perfect Negative Correlation

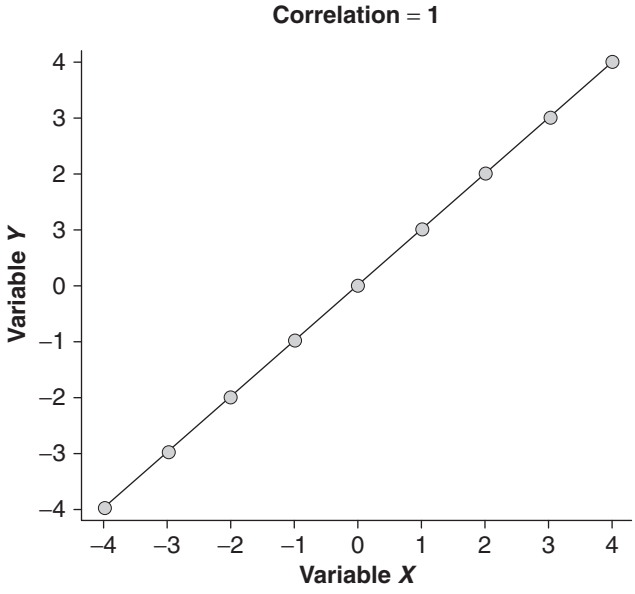



**Figure 8** Perfect Positive Correlation

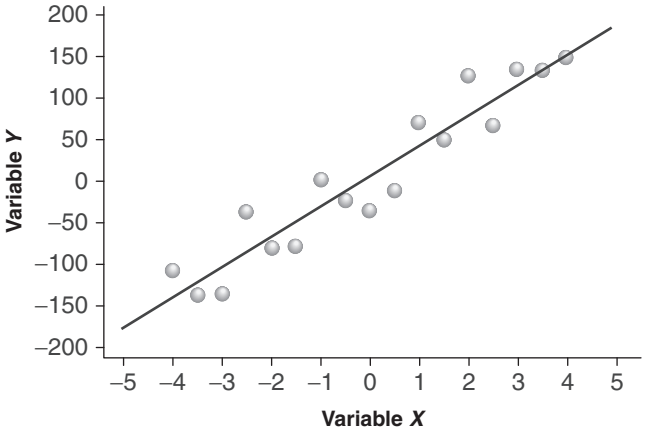



**Figure 10** Scatterplot and Corresponding Regression Line

method is used to get a regression line, which minimizes the sum of the squared residuals. Generally, the line of best fit is also referred to as the regression line.

Before a regression analysis is implemented, a scatterplot is helpful in judging whether the regression line is appropriate to depict the relationship between two variables. Standard statistical software, like SAS, SPSS, R, and S-Plus, can generate a scatterplot with the regression line. Figure 10 shows a scatterplot and the corresponding regression line. This graphic representation indicates it makes sense to use the score from variable *X* to predict the score in variable *Y*.

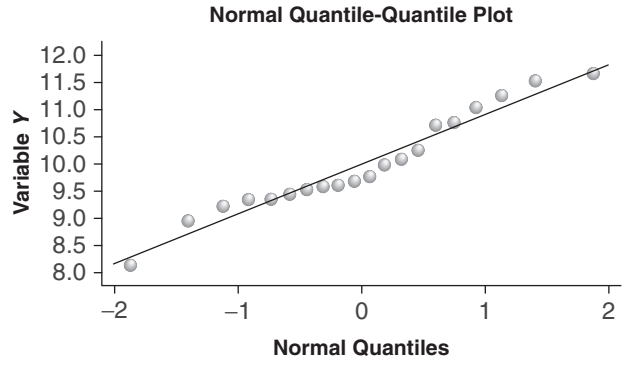



**Figure 11** Q-Q Plot

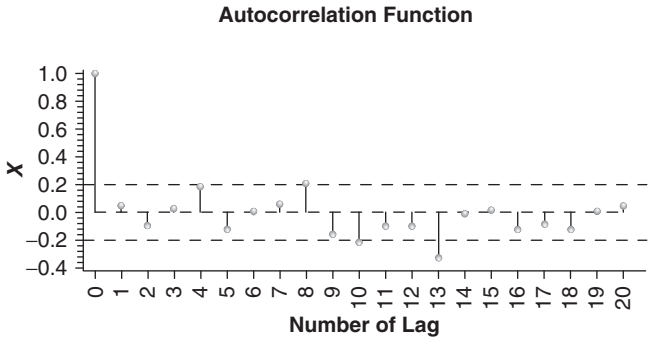



**Figure 12** Stationary Series

**Table 1** Commonly Encountered Departures From Normality

| *Description of Point Pattern* | *Possible Interpretation* |
|---|---|
| Some points do not fall on a line | Outliers in the data |
| Left end of pattern is below the line; right end of pattern is above the line | Long tails at both ends of the data distribution (platykurtosis) |
| Left end of pattern is above the line; right end of pattern is below the line | Short tails at both ends of the distribution (leptokurtosis) |
| Curved pattern with slope increasing from left to right | Data distribution is skewed to the right |
| Curved pattern with slope decreasing from left to right | Data distribution is skewed to the left |
| Staircase pattern (plateaus and gaps) | Data have been rounded or are discrete |

### Test for Normality

Normality assumption is required by most statistical procedures. Statisticians usually use a normal quantile-quantile (Q-Q) plot to check the normality of a variable. Figure 11 is the example of a Q-Q Plot.

The Q-Q plot is a type of bivariate plot that can be used to identify commonly encountered departures from normality as shown in the Table 1.

### Time-Series Analysis

Scatterplots are also used in time-series analysis. It is assumed that the noise series for an autoregressive-moving average model must be stationary, which means that both the expected values of the series and its autocovariance function are independent of time. These scatterplots with needles are called autocorrelation functions (ACFs) because they show the degree of correlation with past values of the series as a function of the number of periods in the past (that is, the lag) at which the correlation is computed. A scatterplot of this series and its autocorrelation function is the standard way to check for nonstationarity. The autocorrelation plot shows how values of the series are correlated with past values of the series. A series is considered nonstationary if a visual inspection of the autocorrelation function plot indicates that the ACF decays very slowly. If the series is nonstationary, then the next step is to transform it to a stationary series by differencing. Figure 12 shows an example of a stationary series because the ACF decays very drastically. Actually, this series was generated by a random number generator.

*Fei Gu and Neal Kingston*

***See also*** Bivariate Regression; Correlation; Kurtosis; Least Squares, Methods of; Line Graph; Normality Assumption; Pearson Product-Moment Correlation Coefficient

**Further Readings**

Myatt, G. J. (2007). *Making sense of data: A practical guide to exploratory data analysis and data mining.* Hoboken, NJ: Wiley.

Tufte, E. R. (2001). *The visual display of quantitative information* (2nd ed.). Cheshire, CT: Graphics Press.

# Scheffé Test

The Scheffé test is one of the oldest multiple comparison procedures in use today. It is important to recognize that it is a frequently misused procedure and that it is also a valuable test when used as Henry Scheffé intended it. Unlike competitors such as Tukey's Honestly Significant Difference test, the Scheffé test is specifically designed for the situation in which post hoc comparisons involve more than pairwise differences. For example, it could be used to compare the mean of two groups to the mean of two other groups on the basis of interesting differences that appeared after the data had been collected.

This entry begins by describing the background and development of the Scheffé test. Next, this entry details the test statistic, its role in variance heterogeneity, and its relation with the Bonferroni correction and a priori tests. Last, the entry discusses how the Scheffé test is implemented in computer programs and its future in the field.

## Background

Like much in statistics, it all started with John Tukey. In a series of oral presentations beginning in about 1950 and culminating in what might be the most cited unpublished manuscript in statistics, Tukey developed the concept of error rates, focusing on the per comparison and familywise rates. Scheffé developed his procedure at the time that Tukey was speaking about his own ideas on multiplicity and, in a footnote to his paper, Scheffé explicitly gave credit to Tukey for the familywise error rate concept that was the basis for his test. The reason for focusing on the familywise error rate as the basis for a test is that it holds the error rate constant at $\alpha$ for the set of all possible contrasts, whether those contrasts had been planned a priori or whether they were reached after examining the data. Scheffé's test was intended specifically as a post hoc test and is most appropriately used that way. It is also a simultaneous inference procedure because intervals or tests are computed simultaneously rather than in a stepwise or layered fashion.

## The Test Statistic

Unlike most multiple comparison procedures, Scheffé's test was built on the standard $F$ distribution. Although his derivation of the method is complex, a very clear explanation can be found in the text by Scott Maxwell and Harold Delaney. The basic idea is to find the distribution of the $F_{\text{maximum}}$ statistic, which is the maximum possible $F$ for *any* contrast on a set of means. Maxwell and Delaney show that the distribution of $F_{\text{maximum}}$ is

$$\begin{aligned} F_{\text{maximum}} &= (k-1)MS_{\text{Bet}}/MS_{\text{error}} \\ &= (k-1)F_{.05;\,k-1,N-k} \end{aligned},$$

where $k$ is the number of groups, $MS_{\text{Bet}}$ and $MS_{\text{error}}$ are the between groups and error mean squares, and $N$ is the total number of observations. The distribution of all possible contrasts (pairwise or complex) will have an $F_{\text{maximum}}$ distribution under the null hypothesis and can be tested against $(k-1)F_{.05;\,k-1,\,N-k}$. Using that critical value, the familywise error rate will be, at most, $\alpha$. If the omnibus null hypothesis is rejected, then at least one contrast will equal $F_{\text{maximum}}$. Although the Scheffé test is not a protected test like Fisher's least significant difference test, which requires a significant omnibus $F$ before any comparisons are made, there is no point in using Scheffé's test unless the omnibus $F$ is significant. If it is not significant, there can be no significant contrasts. In contrast, just because a significant omnibus $F$ indicates that there will be at least one significant contrast, it does not guarantee that the contrast will be of any interest to researchers.

Scheffé's test has been controversial from the beginning. Scheffé and Tukey long debated the relative merits of $F$ and Studentized range statistics as the basis for statistical decisions, and the lower power of Scheffé's test for all pairwise contrasts is well known. The Studentized range statistic is the difference between the largest and smallest means in a set divided by the square root of $MS_{\text{error}}/n$,

where $n$ is the size of any one sample. This statistic has its own distribution that is different from the standard $F$ distribution. In a footnote to his original paper, Scheffé specifically said that if one were only interested in pairwise comparisons between all means, the Tukey test should be used instead because the confidence intervals would be shorter. The advantage of Scheffé's test is that it allows researchers to examine any contrast, whether it was arrived at before the data were collected or after interesting mean differences appeared in the data. Currently, Scheffé's test is the only post hoc procedure that allows one to examine complex contrasts.

As an example of a complex contrast and its confidence interval, suppose that a study was conducted comparing five different forms of therapy and two of them seemed to be considerably more effective than the other three. It would be sensible to wish to compare the mean for those two treatments against the mean of the other three. The means are [210, 205, 190, 180, 185], with $MS_{\text{error}} = 264$ and the size of each sample ($n_j$) is 5. The overall $F$ is 3.116, which is significant at $\alpha = .05$. The contrast of the first two groups against the last three groups is denoted as

$$\psi = \left(\Sigma c_j \overline{X}_j\right),$$

where $c_j = [1/2, 1/2, -1/3, -1/3, -1/3]$. (The simplest way for the researcher to arrive at the necessary coefficients is to break the design into the two subsets he or she wishes to compare, and then to take as the coefficients for members of each subset the reciprocals of the number of groups in that subset, making one subset negative.)

The $F$ for this contrast would be

$$\begin{aligned}\psi = \left(\Sigma c_j \overline{X}_j\right) &= \frac{1}{2}(210) + \frac{1}{2}(205) - \frac{1}{3}(188) \\ &\quad - \frac{1}{3}(187) - \frac{1}{3}(180) = 22.5, \\ F = \frac{n\psi^2}{\Sigma c_j^2 MS_{\text{error}}} &= \frac{n\left(\Sigma c_j \overline{X}_j\right)^2}{\Sigma c_j^2 MS_{\text{error}}} \\ &= \frac{5(22.5)^2}{.833 \times 264} = 11.51.\end{aligned}$$

The critical value for this or any similarly calculated contrast would be $(k - 1)F_{.05;\ k-1,\ N-k} = (4)2.87 = 11.47$, which would cause the rejection of the null hypothesis and the conclusion that the first two therapy conditions are significantly better than the other three conditions. To learn more about this contrast, confidence limits on the mean difference between the two sets of treatments can be calculated.

$$\begin{aligned}\text{CI}_{.95} &= \psi \pm \sqrt{(k-1)F_{.05;\,k-1,\,df}}\sqrt{\frac{\Sigma c_j^2 MS_{\text{error}}}{n}} \\ &= 22.5 \pm \sqrt{11.466}\sqrt{\frac{.833 \times 264}{5}} \\ &= 22.5 \pm (3.386)(6.633) \\ &= 22.5 \pm 22.466 \\ &= 44.966 - 0.034.\end{aligned}$$

Although both ends of the confidence interval are positive, it is a very wide interval with a lower end very close to 0.00. Any number of intervals can be formed in this way and the probability that all intervals would simultaneously include the corresponding parameter is $\alpha$.

The post hoc contrast that was just tested, and its associated confidence interval, could not be tested by other standard multiple comparison techniques in a way that would control the familywise error rate. But two qualifications need to be made for the last statement. First, Tukey would have argued that you would accomplish nearly the same thing, and perhaps something better, by making pairwise comparisons among all five means. However, Tukey's test would show all groups to be a homogeneous set, which is not particularly helpful. The second argument is that the price you pay for being able to run this, or any other, contrast is extremely high relative to the price you would pay for alternative approaches. For example, with five subjects in each of the five groups, Scheffé's test would require an $F$ of

$$(k-1)F_{.05;4,20} = 4 \times 2.866 = 11.46,$$

whereas the comparable value for all pairwise Tukey HSD tests is

$$\begin{aligned}F &= \left(q_{.05;5,20}/\sqrt{2}\right)^2 = (4.23/1.414)^2 \\ &= 2.991^2 = 8.946.\end{aligned}$$

Clearly, the Tukey would be more powerful if one was willing to substitute pairwise comparisons for the contrast that was tested with Scheffé's test.

## Variance Heterogeneity

The Scheffé test, like other tests involving the analysis of variance on means, assumes that error is normally distributed and homogeneous across groups. If this condition is not met, then the original test can be improved by applying what is known as the Brown-Forsythe procedure. In this case, the denominator of $F$ is modified to adjust the estimate of error. The degrees of freedom ($df$) also need to be modified as shown:

$$F = \frac{\left(\Sigma c_j \overline{X}_j\right)^2}{\Sigma\left(c_j^2/n_j\right)s_j^2}$$

and

$$df = \frac{\left(\sum_1^k c_j^2 s_j^2/n_j\right)^2}{\sum_1^k \left(c_j^2 s_j^2/n_j\right)^2/(n_j-1)}.$$

Similarly, the confidence intervals can be defined as

$$\text{CI} = \psi \pm \sqrt{(k-1)F_{.05;k-1,df}}\sqrt{\sum_1^k\left[\left(c_j^2/n_j\right)s_j^2\right]}.$$

Just as with homogeneous variances and equal sample sizes, these intervals will have a probability of $\alpha$ of *simultaneously* including the relevant parameter, which is one of the reasons that they are often called Simultaneous Test Procedures or Simultaneous Inference Procedures.

## The Bonferroni as an Alternative Approach

One alternative that might immediately come to mind would be to apply a Bonferroni correction to evaluate several contrasts that might need to be tested. However, this is not a legitimate alternative when considering post hoc tests. (Using this correction, the researcher simply divides the desired familywise error rate by the number of contrasts run and uses that value of the critical $p$ value for each contrast.) The Scheffé test would allow an infinite number of contrasts to be tested, so as a post hoc procedure, the adjusted $p$ value for the Bonferroni would be $\alpha/\infty$, which is undefined. Certainly that is not going to work. Thus, when making complex post hoc contrasts, holding the familywise error rate at $\alpha$, Scheffé test is going to be the only choice.

## A Priori Tests

It would be possible to apply Scheffé's procedure when several a priori contrasts need to be tested, and there might be situations in which that would be a practical approach. However, it is not likely that there would be enough a priori contrasts to warrant taking this step. For example, consider the study of five different therapies that were just examined. If two of those therapies were control treatments, then it would be reasonable to wish to compare the control treatments with the others using the contrast given previously. It would also be reasonable to compare the two controls with each other and perhaps planning pairwise comparisons among the three active treatments. But that would only come to five contrasts. The researcher might come up with one or two others that he or she would like to test, so perhaps there are seven contrasts. If Scheffé's test was used, the critical value of $F$ would be 11.46, as shown previously. Alternatively, if a Bonferroni correction was applied to the tests, it would require a significance level of $\alpha = .05/7 = .007$, which is equivalent to an $F$ of only 4.807. Certainly, the researcher would rather work against a critical value of 4.807 than one of 11.46. Thus, the Bonferroni correction for the seven a priori contrasts is considerably more powerful than the Scheffé test, which would allow many more contrasts to made that are not likely to be of interest. In fact, the researcher would have to run 18 contrasts for the Scheffé test to be more powerful than the Bonferroni in this situation, and it would be difficult to imagine any way the researcher could think of 18 sensible contrasts among five means.

## Implementing Scheffé's Test in SAS and SPSS

Both SAS and SPSS, an IBM product, offer the Scheffé test as a post hoc test but only for

pairwise comparisons. As noted earlier, the Scheffé should not be used for only pairwise tests. However, both programs allow the user to specify contrasts, although the user will have to calculate the critical value himself or herself, which is hardly a problem. SPSS allows the user to specify contrasts for the "One-way ANOVA" procedure, and SAS allows the user to define contrasts under "Proc GLM." Neither program will print out confidence intervals, so the user needs to do that by hand.

### The Changing Landscape

When Tukey started working on multiple comparison procedures, he largely set the field moving in the direction of familywise error rates, which is where Scheffé's test fits. But the problem with the familywise error rate approach is that it is an extreme case. If a family consists of a large number of contrasts (pairwise or not), then any error is considered a failure of the entire set—the confidence intervals no longer cover *all* parameters. That is why the critical values are as large as they are. In the last decade of Tukey's life, he began to take an interest in Benjamini's False Discovery Rate, which aims to control the *number* of false discoveries (Type I errors) that a procedure will make, rather than focusing on allowing no errors. As the field moves further in that direction, tests such as Scheffé's, and perhaps even Tukey's, are likely to play a smaller role.

*David C. Howell*

***See also*** A Priori Monte Carlo Simulation; Bonferroni Procedure; *F* Test; Multiple Comparison Tests; Null Hypothesis; Omnibus Tests; Post Hoc Comparisons; Tukey's Honestly Significant Difference (HSD)

#### Further Readings

Howell, D. C. (2010). *Statistical methods for psychology* (7th ed.). Belmont, CA: Thomson/Wadsworth.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Scheffé, H. (1953). A method for judging all contrasts in the analysis of variance. *Biometrika, 40,* 87–104.

## Scientific Method

The term *method* derives from the Greek *meta* and *odos* meaning *following after,* suggesting the idea of order. Applied to science, method suggests the efficient, systematic ordering of inquiry. Scientific method, then, describes a sequence of actions that constitute a strategy to achieve one or more research goals. Relatedly, scientific *methodology* denotes the general study of scientific methods and forms the basis for a proper understanding of those methods.

Modern science is a multifaceted endeavor. A full appreciation of its nature needs to consider the aims it pursues, the theories it produces, the methods it employs, and the institutions in which it is embedded. Although all these features are integral to science, science is most illuminatingly characterized as method. Method is central to science because much of what we have learned from science has been acquired through use of its methods. Our scientific methods have been acquired in the course of learning about the world; as we learn, we use methods and theorize about them with increased understanding and success.

In this entry, scientific method is contrasted with other types of method. Then, some criticisms of the idea of scientific method are considered. Thereafter, four major theories of scientific method are outlined and evaluated. Finally, the place of methods in science is addressed.

### Four Methods of Knowing

The American philosopher and scientist Charles Sanders Peirce maintained that there are four general ways of establishing beliefs. The poorest of these, the *method of tenacity,* involves a person stubbornly clinging to a familiar idea when challenged. The belief is sustained by an attitude of tenacity and unquestioned acceptance. The *method of authority* maintains that ideas are held to be true simply because they are the ideas of a person who is deemed an expert or perceived to be in a position of power. Peirce noted that this method is superior to the method of tenacity, because some beliefs can be fixed by adopting the method. The *a priori method,* which is better than both of the methods just mentioned, involves an

appeal to the powers of reason independent of scientific observation. It involves accepting beliefs on the grounds that they are intuitive, self-evident, and based on reason rather than experience. The *method of science* is the method that Peirce himself advocated. It is superior to the other three methods because it establishes belief by appeal to an external reality and not to something merely human. Unlike the other methods, which are prescientific, the method of science has the characteristic of self-correction because it has built-in checks along the way. For Peirce, only this method has the ability to lead eventually to the truth.

## Criticisms of the Idea of Scientific Method

Despite the importance of method in science, the idea that there is a scientific method characteristic of scientific inquiry has been the subject of many criticisms. Perhaps the most frequently voiced criticism of scientific method is that there is no such thing as *the* scientific method; that is, there can be no fixed universal account of scientific method appropriate for all disciplines and at all times. This criticism should be readily accepted because it speaks against an unrealistic view of scientific method.

Another prominent criticism of scientific method was proposed by Karl Popper, who often remarked that scientific method does not exist. By this he meant that there is no method of discovering a scientific theory, that there is no method of verification, and that there is no method for establishing whether a hypothesis is probably true or not. However, these claims are part of Popper's preference for a falsificationist construal of the hypothetico-deductive method. These claims might, or might not, be part of other conceptions of scientific method. For example, advocates of an inductive conception of scientific method will not accept the first claim; those who accept the idea of confirmation, as distinct from falsification, will argue against the second claim; and Bayesian methodologists will reject the third claim.

In a book, which was provocatively entitled *Against Method,* Paul Feyerabend presented a different criticism of scientific method. He argued that there are no methodological rules that are part of scientific method that have not been broken at some time or other in the interests of genuine scientific progress. Thus, for Feyerabend, the only rule that does not inhibit progress is the rule "anything goes." Feyerabend's argument has been endorsed by several commentators who are critical of appeals to the importance of scientific method. However, it should be noted that Feyerabend's criticism strictly speaks against the fixity of methodological rules only. There is nothing in Feyerabend's writing that would counsel against the flexible use of a variety of different methodological rules that are revisable in the light of experience, reason, and other sources of justification.

None of the three criticisms just considered addresses contemporary issues in scientific methodology.

## Four Theories of Scientific Method

Modern scientific methodology has given considerable attention to the following four general theories of scientific method: (1) inductive method, (2) hypothetico-deductive method, (3) Bayesian hypothesis testing, and (4) inference to the best explanation.

### *Inductive Method*

The idea that scientific method involves inductive reasoning goes back at least to Aristotle, and was given a heavy emphasis by Francis Bacon and John Stuart Mill. Inductive reasoning takes different forms. For example, it is to be found in the fashioning of statistical generalizations, in the Bayesian assignment of probabilities to hypotheses, and in the reasoning involved in moving from data to hypotheses in the hypothetico-deductive method.

The most popular inductive approach to scientific method is sometimes called *naïve inductivism.* According to this account of method, science begins by securing observed facts, which are collected in a theory-free manner. These facts provide a firm base from which the scientist reasons upward to hypotheses, laws, or theories. The reasoning involved takes the form of enumerative induction and proceeds in accordance with some governing principle of inductive reasoning. As its name suggests, enumerative induction is a form of argument in which the premises enumerate several observed cases from which a conclusion is drawn, typically in the form of an empirical generalization. However, enumerative induction can also

take the form of a prediction about something in the future or a retrodiction about something in the past. The governing principle for an enumerative induction to a generalization can be stated informally as follows: "If a proportion of As have been observed under appropriate conditions to possess property B, then infer the same proportion of all As have property B." This inductive principle can be taken to underwrite the establishment of statistical generalizations.

The naïve inductive method has been criticized in various ways, although the criticisms are mostly directed at extreme versions of the method—versions making the claim that observed facts can be known infallibly, that observations are made in an entirely theory-free manner, or that empirical generalizations can be secured through the use of a strongly justified principle of induction. However, the so-called naïve inductive method can be defended in a moderate form: Observed facts can be established reliably, if fallibly; theory has to be used to guide observations, and theoretical terms can be used to report observational statements without threatening the reliability of those statements; and principles of induction can be given an adequate justification on pragmatic grounds.

In the behavioral sciences, the radical behaviorism of Burrhus F. Skinner is a prominent example of a research tradition that makes use of a nonstatistical inductive conception of scientific method. The major goals of radical behaviorist research are first to detect empirical generalizations about learning, and then to systematize those empirical generalizations by assembling them into nonexplanatory theories. Murray Sidman's *Tactics of Scientific Research* is an instructive radical behaviorist account of the methodology of phenomena detection.

The Bayesian approach to hypothesis testing can be regarded as a sophisticated variant of inductive method. It is considered later as an account of scientific method in its own right.

### *Hypothetico-Deductive Method*

The most popular account of method in science is the hypothetico-deductive method, which has been the method of choice in the natural sciences for more than 150 years. The method has come to assume hegemonic status in the behavioral sciences, which have often placed a heavy emphasis on testing hypotheses in terms of their predictive success. Relatedly, the use of traditional statistical significance test procedures is often embedded in a hypothetico-deductive structure.

The hypothetico-deductive method is characteristically described in one of two ways: On one account, the scientist takes a hypothesis or a theory and tests it indirectly by deriving from it one or more observational predictions, which are amenable to direct empirical test. If the predictions are borne out by the data, then that result is taken as a confirming instance of the theory in question. If the predictions fail to square with the data, then that fact counts as a disconfirming instance of the theory. The second account is from Karl Popper, who construes the hypothetico-deductive method in falsificationist terms. On this rendition, hypotheses are viewed as bold conjectures, which the scientist submits to strong criticism with a view to overthrowing or refuting them. Hypotheses that successfully withstand such criticism are said to be corroborated, which is a noninductive notion of support.

Even though the hypothetico-deductive method is used by many scientists and has been endorsed by prominent philosophers of science, it has received considerable criticism. Leaving aside Popper's less influential falsificationist account of the hypothetico-deductive method, the major criticism of the hypothetico-deductive method is that it is confirmationally lax. This laxity arises from the fact that any positive confirming instance of a hypothesis obtained through its use can confirm any hypothesis that is conjoined with the test hypothesis, irrespective of the plausibility of that conjunct. Another criticism of the hypothetico-deductive method is that it submits a single hypothesis to critical evaluation without regard for its performance in relation to possible competing hypotheses. Yet another criticism of the method is that it mistakenly maintains that hypotheses and theories arise through free use of the imagination, not by some rational, methodological, or logical means.

Criticisms such as these have led some methodologists to recommend that the hypothetico-deductive method should be abandoned. Although this might be a reasonable recommendation about the method as it is standardly conceived, it is possible to correct

for these deficiencies and use the method to good effect in hypothesis testing research. For example, one might overcome the confirmational defects of the orthodox hypothetico-deductive method by employing a Bayesian approach to confirmation within a hypothetico-deductive framework. With or without a commitment to the Bayesian approach, one could use the hypothetico-deductive method to test two or more competing hypotheses deliberately in relation to the evidence, rather than one hypothesis in relation to the evidence. Furthermore, in testing two or more hypotheses, one might supplement the appeal to empirical adequacy by invoking criteria to do with explanatory goodness.

### *Bayesian Method*

Although the Bayesian approach to evaluating scientific hypotheses and theories is looked on more favorably in philosophy of science than the hypothetico-deductive alternative is, it remains a minority practice in the behavioral sciences.

For the Bayesian approach, probabilities are considered central to scientific hypothesis and theory choice. It is claimed that they are best provided by probability theory, which is augmented by the allied philosophy of science known as *Bayesianism*. In using probability theory to characterize theory evaluation, Bayesians recommend the assignment of posterior probabilities to scientific hypotheses and theories in the light of relevant evidence. Bayesian hypothesis choice involves selecting from competing hypotheses the one with the highest posterior probability, given the evidence. The vehicle through which this process is conducted is Bayes' theorem. This theorem can be written in a simple form as: $Pr(H/D) = Pr(H) \times Pr(D/H) \div Pr(D)$. The theorem says that the posterior probability of the hypothesis is obtained by multiplying the prior probability of the hypothesis by the probability of the data, given the hypothesis (the likelihood), and dividing the product by the prior probability of the data.

Although Bayes' theorem is not controversial as a mathematical theorem, it is controversial as a guide to scientific inference. With respect to theory appraisal, one frequently mentioned problem for Bayesians is that the probabilistic information required for their calculations on many scientific hypotheses and theories cannot be obtained. It is difficult to know how one would obtain credible estimates of the prior probabilities of the various hypotheses and evidence statements that comprised, say, Charles Darwin's evolutionary theory. Not only are the required probabilistic estimates for such theories hard to come by, but also they do not seem to be particularly relevant when appraising such explanatory theories.

The problem for Bayesianism presented by scientific theory evaluation is that scientists naturally appeal to qualitative theoretical criteria rather than probabilities. It will be described in the next section that scientific theories are often evaluated qualitatively by employing explanatory reasoning rather than probabilistic reasoning.

### *Inference to the Best Explanation*

In accordance with its name, inference to the best explanation (IBE) is founded on the belief that much of what we know about the world is based on considerations of explanatory worth. In contrast to the hypothetico-deductive method, IBE takes the relation between theory and evidence to be one of explanation, not logical entailment, and by contrast with the Bayesian approach, it takes theory evaluation to be a qualitative exercise that focuses explicitly on explanatory criteria, not a quantitative undertaking in which one assigns probabilities to theories. Given that a primary function of many theories in science is to explain, it stands to reason that the explanatory merits of explanatory theories should count in their favor, whereas their explanatory failings should detract from their worth as theories. The major point of IBE is that the theory judged to be the best explanation is taken as the theory most likely to be correct. There is, then, a two-fold justification for employing IBE when evaluating explanatory theories: It explicitly assesses such theories in terms of the important goal of explanatory power, and it provides some guide to the approximate truth of theories.

The cognitive scientist Paul Thagard has developed a detailed account of IBE as a scientific method—one that helps a researcher to appraise competing theories reliably through the coordinated use of several criteria. This method is known

as the *theory of explanatory coherence*. The theory comprises an account of explanatory coherence in terms of many constituent principles, a computer program for implementing the principles, and various simulation studies that demonstrate its promise as a method of IBE.

According to the theory of explanatory coherence, IBE is centrally concerned with establishing relations of explanatory coherence. To infer that a theory is the best explanation is to judge it as more explanatorily coherent than its rivals. The theory of explanatory coherence is not a general theory of coherence that subsumes different forms of coherence such as logical and probabilistic coherence. Rather, it is a theory of *explanatory* coherence, where the propositions hold together because of their explanatory relations.

Relations of explanatory coherence are established through the operation of seven principles. These principles are symmetry, explanation, analogy, data priority, contradiction, competition, and acceptability. The determination of the explanatory coherence of a theory is made in terms of the three criteria. Within the theory of explanatory coherence, each of these criteria is embedded in one or more of the seven principles.

Thagard determined that explanatory breadth is the most important criterion for choosing the best explanation. This criterion captures the idea that a theory is more explanatorily powerful than its rivals if it explains a greater range of facts.

The notion of simplicity that Thagard deems most appropriate for theory choice is a pragmatic criterion that is closely related to explanation; it is captured by the idea that preference should be given to theories that make fewer special or ad hoc assumptions. Thagard regards simplicity as the most important constraint on explanatory breadth; one should not sacrifice simplicity through ad hoc adjustments to a theory to enhance its consilience.

Finally, Thagard found that analogy is an important criterion of IBE because it can improve the explanation offered by a theory. Explanations are judged more coherent if they are supported by analogy to theories that scientists already find credible.

The four theories of scientific method just considered are commonly regarded as the major theories of scientific method. Although all the methods have sometimes been proposed as the principal claimant for the title of *the* scientific method, they are better thought of as restrictive accounts of method that can be used to meet specific research goals, not broad accounts of method that pursue a range of research goals.

## The Importance of Method

Even though methodological discussions of scientific method are not fashionable, they are of vital importance to the well-being of science. For it is to scientific methods that scientists naturally turn for the cognitive assistance they need to investigate their subject matters successfully. The evolution and understanding of scientific methods is to be found in the domain of scientific methodology; this fact makes this interdisciplinary sphere of learning of major practical and educational importance.

*Brian D. Haig*

***See also*** Alternative Hypotheses; Bayes's Theorem; Falsifiability; Inference: Deductive and Inductive; *Logic of Scientific Discovery, The*

### Further Readings

Haig, B. D. (2005). An abductive theory of scientific method. *Psychological Methods, 10,* 371–388.

Howson, C., & Urbach, P. (2006). *Scientific reasoning: The Bayesian approach* (3rd ed.). LaSalle, IL: Open Court.

Laudan, L. (1981). *Science and hypothesis: Historical essays on scientific methodology.* Dordrecht, the Netherlands: Reidel.

Nickles, T. (1987). Twixt method and madness. In N. J. Nersessian (Ed.), *The process of science* (pp. 41–67). Dordrecht, the Netherlands: Martinus Nijhoff.

Nola, R., & Sanky, H. (2007). *Theories of scientific method.* Montréal, Québec, Canada: McGill-Queens University Press.

Rozeboom, W. W. (1999). Good science is aductive, not hypothetico-deductive. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 335–391). Hillsdale, NJ: Lawrence Erlbaum.

Sidman, M. (1960). *Tactics of scientific research.* New York: Basic Books.

Thagard, P. (1992). *Conceptual revolutions.* Princeton, NJ: Princeton University Press.

## Secondary Data Source

The analysis of secondary data plays a vital role in many fields of study, including the social sciences. The definition of secondary versus primary data is not based on specific qualities of the data itself but on its history and relationship to a specific analysis. A simple definition is that primary data are collected by a research group for the specific analysis in question, whereas secondary data are collected by someone else for some other purpose. So if a researcher conducts a survey and analyzes the results for his or her analysis, the data from the survey are primary data. If the researcher deposits the data in an archive and someone else unrelated to the original research team analyzes it 20 years later, then the results for that analysis the data are now secondary data.

One reason analysis of secondary data is becoming more popular in the social sciences is the availability of large data sets collected and processed by the government and made available for researchers to analyze. Examples of such data sets in the United States include those from the decennial U.S. Census, which aims to collect information from every person living in the United States in the year the census is conducted, and the annual Behavior Risk Factor Surveillance Survey System (BRFSS), which collects data on health behaviors from a representative sample of Americans and is weighted to reflect the entire U.S. population. It would be beyond the capability of most if not all research teams to collect data on this scale, but the data from these projects are available for anyone with a connection to the Internet to download for free.

The distinction between primary and secondary data should not be overemphasized. Many researchers work with both primary and secondary data during the course of their careers, depending on the specific research questions they are studying at the time, and often both primary and secondary data are analyzed within one research project. The same statistical techniques might be used on either primary or secondary data, and both types of data have advantages and disadvantages. The goal of the researcher, therefore, should be to select appropriate data for a specific research question.

The primary advantage of using secondary data is the fact that the data are already collected and processed, which represents a substantial savings of time and money and allows researchers to focus their efforts on framing questions and conducting analyses. Another major advantage is the scope of secondary data available: Few researchers would be able to conduct even one survey comparable with the BRFSS, for instance, which has been conducted annually since 1984. A third advantage is that the quality of secondary data is often very high: For instance, federal agencies have a large staff that trained to plan large-scale surveys, write data collection instruments, conduct surveys, and clean the data. Such projects often use scientific sampling plans which allow for data to be weighted to represent larger populations, such as the entire U.S. population. The methodologies applied to the data are often well documented as well: For instance, the Centers for Disease Control and Prevention, which is a branch of the U.S. National Institutes of Health, issues many technical reports (which are freely available on the Internet) describing their data sets and the methods by which the data were collected and processed.

The major disadvantage of using secondary data is that a given data set might not suit a particular research question: It is a case of having to use what is available rather than deciding what data would be best suited for a desired analysis. With secondary data, there is almost inevitably a gap between the analyst's interests and the purposes for which the data were originally collected, and sometimes researchers take the opposite tack and shape their research questions according to the availability of secondary data. Although this has the benefit of allowing more time to be spent framing questions and conducting analyses, it is contrary to the usual way of conducting research, where one begins by deciding what questions are most interesting or urgent, and then shapes the research enterprise to seek answers to them.

A second problem researchers confront when working with secondary data is that important ancillary information might not be available. For instance, what was the original research question? What was the sampling plan? How well was it executed? Was the survey revised after the study began? What was the response rate? How were the data cleaned after collection? What was done

about missing data? Were any transformation or imputation processes applied? In the case of federally sponsored research projects, this information is usually readily available, but that might not be the case for data from smaller projects conducted by individual research laboratories. Without sufficient information about the original research design and the data collection and cleaning process, the researcher cannot be sure what the data collected actually represents or what methods of analysis are appropriate.

The choice of primary or secondary data is not always an either/or question: Instead, efforts should be focused on creating a research plan that is appropriate to examine the questions of interest. When evaluating whether secondary data will be useful in the research enterprise, the following major issues should be considered:

- Are data available for the time periods I want to study? Even if a survey is conducted annually, often there is a lag of a year or more before data are available for analysis.
- Are data available for the geographic regions I want to study?
- Are data available at an appropriate level of detail for my research plan? This is an important consideration because often data are available only in aggregated format for confidentiality reasons, even if it was collected at a more granular level.
- Are data available for the major variables in my research design, in a form that is useful to me? This is particularly important for data such as race and ethnicity, because many different coding categories have been used over the years, some of which do not correspond to current ways of categorizing this information.
- Is ancillary information about the data available, including sufficient information about the original study design, the data collection process, and any cleaning, transformation, or imputation procedures that might have been applied, that I am confident I can analyze the data appropriately?

*Sarah E. Boslaugh*

***See also*** Cluster Sampling; Convenience Sampling; Data Mining; Nonprobability Sampling; Primary Data Source; Probability Sampling; Stratified Sampling; Threats to Validity

**Further Readings**

Boslaugh, S. (2007). *Secondary data sources for public health: A practical guide.* New York: Cambridge University Press.

Bulmer, M. I. A., Sturgis, P. J., & Allum, N. (2006). *The secondary analysis of survey data.* Thousand Oaks, CA: Sage.

Jiang, Y., Alastair, J. S., & Wild, C. J. (2006). Re-using data from case-control studies. *Statistics in Medicine, 25,* 1377–1389.

Lowrance, W. W. (2003). Learning from experience: Privacy and the secondary use of data in health research. *Journal of Biolaw and Business, 6,* 30–60.

Sieber, J. E. (Ed.). (1991). *Sharing social science data: Advantages and challenges.* Newbury Park, CA: Sage.

# SELECTION

Selection refers to a situation in which data are not representative of the underlying population of interest. In particular, selection occurs when unobserved factors that determine whether an observation is in the data set also help determine the value of the quantity of interest. For example, in a study of willingness to volunteer, it would make little sense to only select those participants who volunteer to participate. An analysis of data that suffer from selection generally produces biased estimates and renders statistical inference problematic. Solutions to the problem commonly involve modeling the selection process and the outcome of interest at the same time to account for how selection influences the observed values.

## Types of Missing Data

Selection generally refers to a particular type of missing data: nonrandom missingness. Also referred to as nonignorable (NI) missingness, it is one of three types of missing data, which include missing at random (MAR) and missing completely at random (MCAR). Data that are MCAR are missing in a way that is completely unrelated to any information in the data set—observations or values of variables are missing with equal probability. Data that are MAR have values that are missing in a way that is related to the observed values of other variables in the data set—the

probability that a value is missing can be predicted with observed variables alone. Data exhibit NI missingness when the value of the missing variable helps explain its missingness above and beyond what can be explained with the information contained in observed variables. Of course, this also means that values of variables included in the data set help determine their inclusion as well.

In its common usage, selection refers to a specific type of NI missingness. A typical situation involves no missing data for all covariates $Z$, but missing values of the quantity of interest $Y$. The values of $Y$ might be missing for a variety of reasons. For example, schools might withhold tests from students who might score relatively low, rendering observation of their scores impossible, or respondents might refuse to answer survey questions about sensitive activities. In general, nonrandom sample selection occurs when observation of $Y$ for an individual (or other unit of analysis) depends on unobserved information that helps determine its value.

The best way to avoid nonrandom sample selection is by careful study design and implementation. In survey contexts, allocating effort on additional follow-ups to avoid unit and item nonresponse can reduce or eliminate selection. Phrasing questions in ways that reduce sensitivity can help as well (perhaps by substituting scales in place of exact responses). In program evaluation, the focus should be on ensuring random assignment and high compliance rates. In real-world situations, however, these remedies might be impossible, often leaving only statistical approaches to correct for selection.

## Consequences

Even though the value of $Y$ might depend on a combination of observed factors, which are measured in $Z$, and unobserved factors, the critical element that determines the consequences of selection is whether observation of $Y$ depends, at least in part, on unobserved factors. Selection that depends only on observed factors does not lead to unrepresentative values of $Y$ given the observed covariates, although if one is interested in the average for the entire population, then one should remember to make adjustments to account for differences in the distribution of covariates between the selected sample and the population of interest.

Selection that does depend on unobserved factors leads to unrepresentative values of $Y$, even for observations with the observed covariates: The value of $Y$ will be systematically too large or too small, so estimates of the average value of $Y$ will be biased. This leads to the most common misperception about sample selection: Researchers often assert that despite the presence of selection, their estimates are representative of the sample being studied, although perhaps not of the population of interest. When the value of the unobserved component is related to the observation of $Y$, then the mean of the unobserved component of $Y$ is not zero among the set of individuals for whom the outcome of interest is observed. Thus, the observed value departs from the expected value, and estimates based on those values are almost always wrong.

When moving from estimation of averages for the quantity of interest to regression models that attempt to relate $Y$ to a set of observed covariates $Z$, additional complications arise. The consequences depend on whether the observed factors included in the model are correlated with the unobserved factors that determine the observation and value of $Y$. When these factors are uncorrelated, then coefficient estimates from a linear regression model will be unbiased (although the intercept term will be biased because the mean of the errors is not zero). When these factors are correlated, then coefficient estimates will be biased.

In the linear regression context, selection can best be understood as a form of omitted variable bias, and it produces similar consequences. To illustrate this, consider that observation of $Y$ depends positively on an observed factor $Z$ and some unobserved random component $u$. When $Z$ is large, $Y$ is likely to be observed for almost any value of the random component, but when $Z$ is small, $Y$ will only be observed for large values of the random component. This induces a negative correlation between $Z$ and $u$ among individuals with observed values of $Y$ because, on average, $u$ must be larger as $Z$ gets smaller to exceed the threshold for observation.

Now consider the relationship between unobserved factors $e$ and observed factors $X$ (possibly including $Z$), which explain $Y$ (for example, $Y = X\hat{\beta} + e$). In the context of selection, $u$ is correlated with $e$, because the unobserved factors that

explain the value of *Y* also help explain whether *Y* is observed. Consider the case of a positive correlation: Because the values of *u* necessary to observe *Y* get larger as *Z* gets smaller, the values of *e* will tend to be larger for the associated observations as well. In the standard linear regression model, one assumes that *e* has mean zero. With selection, this might be true for the population of potential observations, but it is not true among the selected sample because of the correlation between *e* and *u*. Because *Y* is a linear function of its observed and unobserved components, the realized values of *Y* are larger than they should be.

Now assume that *Z* and *X* are related, either because both contain at least one factor in common or because there is a nonzero correlation between at least one variable in each. Because *Z* and *u* are correlated among the selection sample and *u* and *e* are correlated in the context of nonrandom sample selection, a correlation between *Z* and *X* will lead to a correlation between *X* and *e*. If *Z* and *X* are positively correlated, then in the current example, the negative relationship between *Z* and *e* will lead to a negative correlation between *X* and *e*. This violates the regression assumption that the errors are uncorrelated with the explanatory variables, which is also what happens with omitted variable bias and leads to biased estimates of the effect of *X* on *Y*.

## Approaches for Eliminating Selection Bias

The omitted variable bias interpretation was described by James Heckman, who also proposed a solution to the problem. Heckman showed that by accounting for the fact that selection affects the distribution of the unobserved component of *Y* in the observed sample, one can generate accurate regression coefficients and estimates of the conditional expected value of *Y*. The correction assumes that the two errors are distributed bivariate normal, and it is implemented by running a probit model of the selection process then using the estimates to calculate the inverse Mills's ratio (the density divided by the cumulative distribution function evaluated at *Z*) for each individual in the selected sample. The inverse Mills's ratio is then included in the linear regression model as an additional right-hand side variable. This process cleans the errors of the part affected by the selection process and breaks the correlation between *e* and *X*, thereby avoiding bias in the coefficient estimates.

The two-step procedure described previously is unbiased but produces inaccurate standard errors because the uncertainty associated with the model of the selection process, and therefore the inverse Mills's ratio, is ignored. An alternative version of the Heckman estimator avoids this problem by estimating the two equations for selection and the outcome of interest simultaneously.

## Selection Bias in Other Regression Models

The Heckman solution for linear regression has been extended to discrete outcomes for both probit and logit equations of interest, as well as for count and duration outcomes. Not all of these estimators lend themselves to correction through the inclusion of the inverse Mills's ratio in the equation of interest. Rather, one should rely on the simultaneous estimation approach to account for the selection process while analyzing the equation of interest.

A related problem occurs when the sample is truncated rather than censored. When censored, researchers observe everything except the outcome variable *Y;* with truncation, researchers observe nothing for respondents that do not select in. Truncation makes it difficult to estimate the selection equation because there is no variation in selection status. Relying on the distributional assumptions about the errors, however, it is possible to estimate the parameters of the selection process simultaneously with the equation of interest, despite the lack of data on observations that do not select into the sample. These models tend to be difficult to estimate, although several authors have proposed methods to improve estimation.

## Challenges in Selection

Although the previous corrections offer an opportunity to avoid the potential threat to inference caused by nonrandom sample selection, diagnosing and correction for it is not always straightforward. It is necessary to specify correctly not only the equation of interest but also the selection equation. Variables incorrectly omitted from both equations can create selection bias, for example. Even with correct specification, estimating selection models,

especially those with discrete dependent variables, can prove difficult: In many cases the estimation process might fail to converge. Even if convergence is achieved, estimates might have large standard errors; particularly the estimate of the error correlation or the estimate of the correlation can be very close to its bounds.

A related issue arises in the form of identification. If the selection equation and the equation of interest contain exactly the same variables, then identification is achieved through the functional form assumption embodied in the specific distribution chosen for the error terms. Under certain conditions, this can be a fairly weak form of identification. When possible, researchers should include variables that influence selection but not the outcome of interest.

The sequential nature of selection estimators also complicates interpretation of the estimation results. Variables included in both equations influence the outcome variable in two different ways: through the selection process and through the equation of interest. Furthermore, the effect depends on the value of other variables included in the selection equation. Care must be taken to account for both influences when calculating marginal effects or first differences, for example.

*Frederick J. Boehmke*

***See also*** Bias; Bivariate Regression; Multiple Regression

**Further Readings**

Achen, C. H. (1986). *The statistical analysis of quasi-experiments*. Berkeley: University of California Press.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica, 47,* 153–161.

Winship, C., & Mare, R. D. (1992). Models for sample selection bias. *Annual Review of Sociology, 18,* 327–350.

# Semipartial Correlation Coefficient

A squared, semipartial correlation coefficient can be used in connection with multiple regression analysis to measure the strength of the association between the dependent and an independent variable, taking into account the relationships among all the variables. A squared semipartial correlation coefficient is also called a squared part correlation. To illustrate the squared semipartial correlation coefficient, consider data that include final mathematics grades (*MA*), student perception of teacher's academic support (*TAS*) in mathematics class, and positive affect (*PA*) in mathematics class. The sample size is $N = 200$. A multiple regression model for these variables is

$$MA = \alpha + \beta_1 PA + \beta_2 TAS + \varepsilon, \qquad (1)$$

where $\alpha$ denotes the intercept; $\beta_1$ and $\beta_2$ denote the regression coefficients (slopes) for *PA* and *TAS*, respectively; and $\varepsilon$ denotes the residual. The sample squared multiple correlation coefficient for the model is $R_2^2 = .169$, where the subscript 2 indicates that there are two variables in the model and .169 means that 16.9% of the variance in *MA* is associated with the joint variability in *PA* and *TAS*. Suppose the researcher wants to know how much variance in *MA* is uniquely associated with *TAS*. The simple linear regression model

$$MA = \alpha + \beta PA + \varepsilon \qquad (2)$$

has $R_1^2 = .121$ and indicates that 12.1% of the variance in mathematics grades is associated with variability in *PA*. So starting with *PA* in the model, 12.1% of the total variance in mathematics grades is associated with variability in *PA*. When *TAS* is added to the model, 16.9% of the total variance in mathematics grades is associated with joint variability in *PA* and *TAS*. Consequently, 4.8% of the total variance in mathematics grades is uniquely associated with variability in *TAS*. The squared semipartial correlation coefficient for *TAS* is simply

$$\Delta R^2 = R_2^2 - R_1^2 = .169 - .121 = .048,$$

where $\Delta R^2$ is read change in $R^2$ and measures the strength of the association between *MA* and *TAS* taking into account the relationships among all three variables. The squared semipartial correlation coefficient for *PA* can also be computed. The squared correlation coefficient for the model

$$MA = \alpha + \beta TAS + \varepsilon \tag{3}$$

is $R_1^2 = .107$, and the squared semipartial correlation coefficient for *PA* is

$$\Delta R^2 = R_2^2 - R_1^2 = .169 - .107 = .062.$$

When the goal is to determine how strongly *TAS* and *MA* are associated taking into account the relationships among all three variables, the roles of *TAS* and *MA* can be interchanged in Equations 1 and 2:

$$TAS = \alpha + \beta_1 PA + \beta_2 MA + \varepsilon, \tag{4}$$

$$TAS = \alpha + \beta PA + \varepsilon, \tag{5}$$

and $\Delta R^2 = .046$, so 4.6% of the variance in *TAS* is uniquely associated with *MA*. The choice between reporting the 4.8% associated with Equation 1 and the 4.6% associated with Equation 4 should be based on whether *MA* or *TAS* is substantively considered the dependent variable.

A semipartial correlation can also be expressed as a correlation between the dependent variable and a residualized version of an independent variable. Let the predicted value for the regression of *TAS* on *PA* be denoted by $\widehat{TAS}$. The predicted value is

$$\widehat{TAS} = a + bPA,$$

where *a* is the sample intercept and *b* is the sample slope for *PA*. The residual is $TAS - \widehat{TAS}$, which is uncorrelated with *PA* and can be interpreted as the part of *TAS* that is not associated with *PA*. The semipartial correlation between *MA* and *TAS* is the correlation between *MA* and $TAS - \widehat{TAS}$, which is often written as $r_{MA(TAS \cdot PA)}$, and can be interpreted as the correlation between *MA* and the part of TAS that is not associated with *PA*. In a similar fashion, $r_{TAS(MA \cdot PA)}$ can be formulated based on

$$\widehat{MA} = a + bPA$$

and

$$MA - \widehat{MA}.$$

The squared semipartial correlation coefficient can also be used in situations with more than two independent variables. In general, let there be *K* independent variables and let the squared multiple correlation coefficient for the *K* variables be denoted by $R_K^2$, where the *K* subscript indicates a model with *K* independent variables. Now if one of the independent variables is removed from the model, the new model has $K - 1$ independent variables. The squared multiple correlation coefficient for the new model is denoted by $R_{K-1}^2$, and its magnitude will be dependent on which of the *K* independent variables is deleted. The squared semipartial correlation coefficient is $\Delta R^2 = R_K^2 - R_{K-1}^2$ and can be interpreted as the percentage of total variance in the dependent variable that is uniquely associated with the deleted variable.

## Hypothesis Test for $\Delta R^2$

Typically, a researcher would be interested in knowing whether the variance uniquely associated with an independent variable is significantly different from zero. That is, using $\mathbf{R}_K^2$ and $\mathbf{R}_{K-1}^2$ to denote the population squared multiple correlation coefficients for the two models, the researcher wants to test the null hypothesis $H_0 : \mathbf{R}_K^2 - \mathbf{R}_{K-1}^2 = 0$ against the alternative $H_1 : \mathbf{R}_K^2 - \mathbf{R}_{K-1}^2 > 0$. The test statistic for this hypothesis is

$$F = (N - K - 1)\frac{R_K^2 - R_{K-1}^2}{1 - R_K^2}, \tag{6}$$

and the critical value is $F_{1-\alpha,\,1,\,N-K-1}$, where $\alpha$ is the Type I error rate for the test. To illustrate, consider testing the significance of the squared semipartial correlation coefficient for *TAS*, that is, the difference between $R^2$ of the full model given in Equation 1 and the reduced model given in Equation 2. From the results reported earlier,

$$\begin{aligned} F &= (N - K - 1)\frac{R_K^2 - R_{K-1}^2}{1 - R_K^2} \\ &= (200 - 2 - 1)\frac{.169 - .121}{1 - .169} = 11.40 \end{aligned},$$

and the $\alpha = .05$ critical value is $F_{1-\alpha,\,1,\,N-K-1} = F_{1-.05,\,1,\,200-2-1} = 3.89$. The null hypothesis is rejected and there is evidence that the

variance uniquely associated with *TAS* is different from zero. The *F* statistic would also be 11.40 if the *F* statistic in Equation 6 were used in conjunction with the regression models in Equations 4 and 5 to test whether the *TAS* variance that is uniquely associated with variance in *MA* is significantly different from zero. That is, either Equations 1 and 2 or Equations 4 and 5 yield the same conclusion about whether there is an association between *MA* and *TAS,* taking into account the relationships among *MA, TAS,* and *PA*.

Returning to the multiple regression model in Equation 1, the regression coefficient for *TAS* is $\beta_2$. If $\beta_2 = 0$, then it would make sense that $\mathbf{R}_2^2 - \mathbf{R}_1^2$ would be zero for *TAS* and thus that a hypothesis test of $H_0 : \beta_2 = 0$ against $H_1 : \beta_2 \neq 0$ would yield the same conclusion as would a test of $H_0 : \mathbf{R}_2^2 - \mathbf{R}_1^2 = 0$ against the alternative $H_1 : \mathbf{R}_2^2 - \mathbf{R}_1^2 > 0$. In fact this is the case, so that a test on the regression coefficient for a variable can be used as the test on the squared semipartial correlation coefficient for that variable.

## Comparison to the Squared Partial Correlation Coefficient

The squared semipartial correlation coefficient measures the strength of association between the dependent variable and one of the independent variables taking into account the relationships among all the variables. The squared partial correlation coefficient, which is denoted by $pr^2$ for partial *r* squared, can be used to accomplish the same aim. A comparison between $pr^2$ and $\Delta R^2$ is presented next in the general context in which there are *K* independent variables. In the following, the variable that is being correlated with the dependent variable is denoted by $X_j$ and the remaining $K - 1$ variables are referred to as the "other" variables.

The squared semipartial correlation coefficient for $X_j$ is

$$\Delta R^2 = R_K^2 - R_{K-1}^2, \tag{7}$$

and the squared partial correlation coefficient is

$$pr^2 = \frac{R_K^2 - R_{K-1}^2}{1 - R_{K-1}^2}. \tag{8}$$

First, note that numerator of $pr^2$ is equal to $\Delta R^2$. Therefore, both $pr^2$ and $\Delta R^2$ address the question of the amount of variance that is uniquely associated with $X_j$. In fact, if $pr^2$ is zero, then $\Delta R^2$ must be zero and vice versa. Similarly, if $\Delta R^2$ is significantly different from zero by using the *F* statistic in Equation 6, then $pr^2$ must also be significantly different from zero. Second, note that $R_{K-1}^2$ is the proportion of the total variance in the dependent variable that is associated with the other variables. Logically, the proportion of total variance in the dependent variable that is associated with the other variables (i.e., $R_{K-1}^2$) plus the proportion of the total variance in the dependent variable that is not associated with the other variables must add to 1.0. Therefore, $1 - R_{K-1}^2$ is the proportion of the total variance in the dependent variable that is not associated with the other variables. As a result, $pr^2$ addresses the following question: "Of the dependent variable variance *not associated with the other variables,* what proportion is uniquely associated with $X_j$?" By contrast, $\Delta R^2$ addresses the following question: "Of the total dependent variable variance, what proportion is uniquely associated with $X_j$?" Thus, it is clear that both are concerned with the amount of variance uniquely associated with $X_j$, but $\Delta R^2$ expresses the unique variance as a proportion of the total variance for the dependent variable, whereas $pr^2$ expresses the unique variance as a proportion of a part of the total variance for the dependent variable. By comparing Equations 7 and 8, it can be observed that, because $1 - R_{K-1}^2$ must be between zero and one, $pr^2$ must be at least as large as $\Delta R^2$, and therefore $pr^2$ will provide an impression of a stronger relationship between the dependent variable and $X_j$ than does $\Delta R^2$.

Because both $pr^2$ and $\Delta R^2$ address the amount of dependent variable variance that is uniquely associated with $X_j$, a reasonable basis for choosing between these statistics is the ease of interpretation and comparison across independent variables. The coefficient $\Delta R^2$ expressed as part of a total seems easier to understand. And if $\Delta R^2$ is available for several variables in the model, they are easy to compare because each is a proportion of the same total.

The partial correlation can be formulated as a correlation between a residualized dependent variable and a residualized independent variable. Consider the example of *K* independent variables

and interest in the partial correlation between the dependent variable $Y$ and $X_j$ taking into account the relationships among $Y$ and all the independent variables. Let $Y - \widehat{Y}_{K-1}$ be the residual after relating $Y$ to the set of independent variables that does not include $X_j$ and let $X_j - \widehat{X}_{j,K-1}$ be the residual after relating $X_j$ to the set of independent variables that does not include $X_j$.

The partial correlation is the correlation between the two residual variables and can be interpreted as the correlation between the parts of $Y$ and $X_j$ that are not related to the other independent variables, whereas the semipartial correlation is the correlation between $Y$ and the part of $X_j$ that is not related to the other independent variables.

*James Algina and H. J. Keselman*

***See also*** Correlation; Multiple Regression; Partial Correlation

**Further Readings**

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Darlington, R. B. (1990). *Regression and linear models*. New York: McGraw-Hill.

Pedhauzer, E. J. (1997). *Multiple regression in behavioral research* (2nd ed.). Fort Worth, TX: Harcourt Brace College Publishers.

Werts, C. E., & Linn, R. L. (1969). Analyzing school effects: How to use the same data to support different hypotheses. *American Educational Research Journal, 6,* 439–447.

# SENSITIVITY

Sensitivity is one of four related statistics used to describe the accuracy of an instrument for making a dichotomous classification (i.e., positive or negative test outcome). Of these four statistics, sensitivity is defined as the probability of correctly identifying some condition or disease state. For example, sensitivity might be used in medical research to describe that a particular test has 80% probability of detecting anabolic steroid use by an athlete. This entry describes how sensitivity scores are calculated and the role of sensitivity in research design.

## Calculating Sensitivity Scores

Sensitivity is calculated based on the relationship of the following two types of dichotomous outcomes: (1) the outcome of the test, instrument, or battery of procedures and (2) the true state of affairs. The outcome of the test is typically referred to as being positive (indicating the condition is present) or negative (the condition is not present). The true state of affairs is typically defined either by assignment of some experimental condition or classification based on some known gold standard test. Based on these two types of dichotomous outcomes, there are four possible outcome variables, which are defined as follows:

*True positive* = the number of cases with a positive test outcome who *do* have the condition

*True negative* = the number of cases with a negative test outcome who *do not* have the condition

*False positive* = the number of cases with a positive test outcome who *do not* have the condition (Type I error)

*False negative* = the number of cases with a negative test outcome who *do* have the condition (Type II error)

Table 1 shows each of these four variables and the statistics generated from them. Sensitivity is based on the variables in the *Does Have the Condition* column and is calculated as the number of True Positives divided by the number of True Positives plus the number of False Negatives.

When a test is measuring some characteristic on a continuous scale, the sensitivity might be changed depending on the cutoff used to define a positive test. To demonstrate this, consider an example in which 5$\alpha$-dihydrotestosterone (DHT) levels were measured from 100 athletes, 50 of which were administered a dose of anabolic steroid prior to their test. Table 2 shows, from left to right, simulated data showing DHT levels and the corresponding number of athletes receiving

**Table 1** Calculation of Sensitivity and Its Relationship to Other Measures of Test Accuracy

| | | *Truth* | | |
|---|---|---|---|---|
| | | *Does Have the Condition* | *Does Not Have the Condition* | |
| *Dichotomous Test Outcome* | Positive Test | True Positive | False Positive | Positive Predictive Value |
| | Negative Test | False Negative | True Negative | Negative Predictive Value |
| | | *Sensitivity* | *Specificity* | |

steroids who have that level of DHT. If a positive test is defined as being any athlete with a DHT level of at least $10^{-12}$ mol/l, then 49 of the 50 steroid-administered athletes would have a positive test outcome (sample size from the $10^{-12}$, $10^{-11}$, and $10^{-10}$ groups). Given this cutoff then, the sensitivity of the procedure would be calculated as: [49 (true positives) ÷ [49 (true positive) + 1 (false negative)]] = 98%. In contrast, if a positive test is defined as a DHT level of at least $10^{-10}$, then 30 of the 50 steroid-administered athletes would have a positive test outcome resulting in a sensitivity of 60% (30 ÷ [30 + 20]).

The athletes who did not receive the drug are not listed in Table 2 because they would by definition either be "True Negative" or "False Positive" cases, which are not part of the sensitivity calculation. They would instead be pertinent to the calculation of specificity. This example demonstrates how, because sensitivity and specificity are calculated on two different samples of individuals, both of these statistics are free to vary from one another independently.

## Sensitivity in Research Design

Sensitivity is typically used to demonstrate or evaluate the accuracy of a test for correctly identifying some condition or disease state. This information is useful both for developers of new instruments and the consumers who use them. Depending on the context though, high sensitivity might or might not be a pertinent variable for evaluating test accuracy. For instance, high sensitivity is typically desired for tests like screening measures that are designed to identify potential disease. Because these measures are typically followed up with more rigorous testing, the cost of false positives at screening is low because these cases can be identified by other tests with greater specificity later. However, the cost of a false negative during screening is high, because it might result in the condition remaining untreated. Therefore, sensitivity is very important during the screening stage of testing.

**Table 2** Example Showing 5α-Dihydrotestosterone Among Athletes Receiving Steroids

| *5α-dihydrotestosterone (DHT) mol/l* | *Number of Athletes Receiving Steroid* |
|---|---|
| $10^{-13}$ | 1 |
| $10^{-12}$ | 5 |
| $10^{-11}$ | 14 |
| $10^{-10}$ | 30 |

Sensitivity is not the only relevant measure of test accuracy. One way to achieve perfect sensitivity of a test is to have a positive outcome in every case. However, this might result in a high rate of false positives as well, so specificity (the capacity to correctly identify those as not having the condition) is also important. Also, because of the way that sensitivity is defined, it is a good measure of one aspect of a test's accuracy for group classification when the true classification of all individuals is known. However, sensitivity is not the most pertinent measure of test accuracy in clinical settings where decisions are being made about an individual's test outcome and the true diagnoses or classification is unknown. In this setting, what is of greater interest is the confidence in interpreting an individual test outcome, which is defined by positive and negative predictive values (see Table 1). Unlike sensitivity, predictive values take into account population base rate of the phenomenon of interest.

*Charles W. Mathias*

***See also*** Dichotomous Variable; Specificity; True Positive; Type I Error; Type II Error

**Further Readings**

Bianchini, K. J., Mathias, C. W., & Greve, K. W. (2001). Symptom validity testing: A critical review. *The Clinical Neuropsychologist, 15,* 19–45.

Glaros, A. G., & Kline, R. B. (1988). Understanding the accuracy of tests with cutting scores: The sensitivity, specificity, and predictive value model. *Journal of Clinical Psychology, 44,* 1013–1023.

Loong, T.-W. (2003). Understanding sensitivity and specificity with the right side of the brain. *BMJ, 327,* 716–719.

# Sensitivity Analysis

Sensitivity analysis is the study of stimulus and response in a system, often characterized as the apportionment of variation in a system's output among several inputs. The system plays the role of a black box, known only by its inputs and outputs. Sensitivity analysis reveals the effective working of the system by showing which variations of inputs affect the values of the outputs.

Historically, sensitivity analysis emerged as a technique in probabilistic risk analysis, particularly in environmental impact analyses. In this context, deterministic models were executed many times with different parameter sets. The probabilistic nature of the analysis came from selecting parameter values consistent with parameter distributions for the inputs. Consequently, the methods of sensitivity analysis were developed to deal with deterministic computer models where experiments can be reproduced exactly, and both inputs and outputs are known with high precision. These techniques are most effective in this context but also apply to systems with internal sources of variation, such as natural systems or stochastic simulations. A larger design might be required to deal with the resulting noise in the data.

## Role of a Sensitivity Analyst

A sensitivity analyst performs several diverse tasks: determining the range or distribution of each input parameter, devising hypotheses for possible system behavior and suitable experimental designs to test the hypotheses, carrying out experiments and analyzing results.

During the life cycle of a sensitivity analysis, goals change. At first, the analyst might stress test the system (typically a computer implementation of a complex mathematical model) to ensure that it behaves as expected. A suitable hypothesis would be that the system can be evaluated and that it will generate appropriate results, everywhere in the sample space. A traditional two-level fractional factorial design, using combinations of extreme parameter values, is particularly effective at rooting out anomalous behavior, exhibited as abnormal program termination or unrealistic outputs. A resolution IV design is ideal as it requires only $2N$ simulations (where $N$ is the number of parameters) to ensure equal frequencies for all eight combinations of high and low values for any selection of three parameters. Where some parameters have infinite distributions (e.g., the normal distribution), a consistent policy is needed for truncation to yield extreme values.

Next, the analyst might screen parameters, hoping to identify a small group that dominate system variation. An appropriate hypothesis is that no individual parameter has any influence on the output(s) of interest. Influential parameters are those for which this hypothesis can be rejected at a high level of significance. If enough runs are performed that the effects of many parameters can be detected, influential parameters are those with the largest effects on system outputs. Initial stress testing runs might have identified parameters with large main effects. These experiments might also have shown the existence of large two-factor interactions, without uniquely identifying them. These effects, based on extreme simulations, might overwhelm less dramatic influences that play a significant role in more probable simulations. The analyst has several design choices to highlight other influential parameters.

Designs where only the value of one parameter changes at a time are particularly effective in identifying even small effects, provided that the experiments are deterministic. It is typically necessary to change the value of a single parameter several times at different locations in the sample space, as the parameter's influence might depend on the values of other parameters. The elementary effects method provides a scheme for varying one parameter at a time on a grid. For each parameter, this method computes an average of the absolute value of the output changes resulting from changing that

parameter alone, to provide a statistic to compare against other parameters.

One-at-a-time designs are inefficient if the number of parameters is large and the number of influential parameters is small, as most of the experiments will reveal little of interest about the system. A common alternative is the Latin hypercube design, which was originally developed for sensitivity analysis applications. In this design, each parameter is stratified on *M* levels, where *M* divides into the number of experiments. In the simplest approach, the levels are randomly assigned to experiments, independently for each parameter. The stratification gives good coverage for each parameter, but unintended correlations might arise among parameters, particularly where the number of experiments is not much larger than the number of parameters. Several methods have been used to control correlations so that influential parameters can be identified uniquely, including Ronald L. Iman and William J. Conover's original approach and the use of joint Latin hypercube/two-level fractional factorial designs.

After the screening stage, the choice of detailed analysis will depend on the goals of the analysis (e.g., to optimize a particular output) and on the results obtained from screening. It is common to go back and change the system at this point. For instance, an influential part of a model might be chosen for subsequent refinement. Model developers must be careful to avoid a feedback situation where a submodel has an error that makes it appear noninfluential, leading to a low priority for future revisions that would correct the error.

When sensitivity analysis is too expensive in terms of the number of experiments required, analysts sometimes turn to group sampling. In this approach, parameters are collected into groups, and every parameter in the group is treated the same way. Sensitivity analysis then determines the effect of the group as a whole. To determine which parameter is causing a group effect, the analysis can be repeated with a different grouping of the parameters. Logical inference can identify the common parameter in influential groups.

### Relationship to Data Mining

Data mining is a related discipline where existing databases are analyzed to determine system behavior. Data mining is analogous to an observational science like astronomy, in which experiments involve only selection and analysis of accessible data. In contrast, sensitivity analysis is more like physical or biological experimentation, where systems can be manipulated to test precise hypotheses.

*Terry Andres*

***See also*** Data Mining; Estimation; Experimental Design; Factorial Design; Parameters; Probability Sampling; Stratified Sampling; Variability, Measure of

#### Further Readings

Andres, T. H. (1997). Sampling methods and sensitivity analysis for large parameter sets. *Journal of Statistical Computation and Simulation, 57,* 77–110.

Iman, R. L., & Conover, W. J. (1980). Small sample sensitivity analysis techniques for computer models, with an application to risk assessment. *Communications in Statistics—Theory and Methods, 49,* 1749–1874.

Iman, R. L. & Conover, W. J. (1982). A distribution-free approach to inducing rank correlation among input variables. *Communications in Statistics, B11(3),* 311–334.

Saltelli, A., Chan, K., & Scott E. M. (Eds.). (2000). *Sensitivity analysis.* West Sussex, UK: Wiley.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., et al. (2008). *Elementary effects method* in *global sensitivity analysis: The primer* (pp. 109–131). West Sussex, UK: Wiley.

## Sequence Effects

Sequence effects are potential confounding influences in experiments where subjects are exposed to multiple conditions. Sequence effects refer to potential interactions among conditions of an experiment based on the sequences these treatments are presented. Sequence effects are distinct from order effects, where the actual order of conditions influences the outcome, and carryover effects, where subjects are permanently changed by the manipulation. To illustrate the differences among sequence, order, and carryover effects, imagine an experiment where subjects are asked to pick up and guess the weight of different objects

on multiple daily sessions. A sequence effect would be the perceived weight of a given object being influenced by whether a light or heavy object was handled just before. In contrast, an order effect would be the perceived weight of objects increasing as the experimental session progresses and subjects grow fatigued. Finally, a carryover effect would be the perceived weight of objects decreasing across sessions as subjects grow stronger from all the excessive lifting.

### Controlling for Sequence Effects

Multiple methods are available for controlling for sequence effects. Ideally, sequence effects can be controlled for within individual subjects. Failing that, experimenters will need to control for sequence effects across subjects.

#### *Controlling for Sequence Effects Within Subjects*

Sequence effects can be controlled for by counterbalancing experimental conditions within subjects, where each subject is exposed to every possible combination of experimental conditions. This is only practical when there are relatively few experimental conditions. A more restrictive means of counterbalancing experimental conditions within subjects is to use reverse counterbalancing, where experimental conditions are presented first in one order and then again in reverse order. For example, suppose participants are asked to rate the intensity of a bright, a medium, and a dim light stimulus after placebo or a drug treatment suspected to cause mild light sensitivity. The experimenter suspects that viewing the bright light makes the next light viewed seem less intense, and conversely, viewing the dim light makes the next light viewed seem more intense. To control for these influences, the researcher first presents the stimuli in the order of bright, medium, and then dim, and after a break presents the stimuli in the reverse order of dim, medium, and then bright. Then, the researcher will average intensity ratings obtained for each stimuli during the two orders.

#### *Controlling for Sequence Effects Between Subjects*

If counterbalancing within subjects is not possible or not practical for some reason, another strategy is to counterbalance conditions between subjects. Suppose the investigator in the previous example deemed it necessary to test all six possible combinations of the three light stimuli. In this case, it would be possible to present each of the six different orders to individual subjects, match these orders across placebo and drug conditions, and compare the average stimuli intensity ratings after placebo and drug treatment. In experiments examining more conditions, it becomes more practical to control for sequence effects using a balanced Latin square. In a balanced Latin square, each condition is immediately preceded once by all other conditions. A balanced Latin square cannot control for sequence effects as completely as counterbalancing; however, it does offer a reasonable degree of control with far fewer subjects.

*Ashley Acheson*

***See also*** Latin Square Design; Order Effects; Single-Subject Design; Within-Subjects Design

#### Further Readings

McBurney, D. H., & White, T. L. (2007). *Research methods* (7th ed.). Belmont, CA: Thomson Higher Education.

Thomas, J. C., & Hersen, M. (2003). *Understanding research in clinical and counseling psychology* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

## Sequential Analysis

During the past decade, various sequential design methods in clinical trial designs have been developed to allow for interim decisions and trial modifications based on cumulated information. As part of well-known adaptive designs in group sequential analysis, Yu Shen and Lloyd Fisher proposed a sequential strategy for monitoring clinical trials, namely self-designing trials. The final test statistic is a weighted average of the sufficient statistic from each stage, where the weight for each stage is determined by the observed data prior to that stage. One important feature of this self-designing trial is that the maximum sample size or the maximum number

of interim analyses, $m$, is not specified in advance but at a random stopping time. The flexibility of modifying the sample size allows one to save an underpowered study when unexpected events occur.

## Preliminaries of a Self-Designing Trial

Consider an equally randomized clinical trial to compare a treatment, $T$, and a control, $C$. Assume that the difference between the two arms is normally distributed with an unknown mean $\theta$ and known variance $\sigma^2$. The methodology can be extended to the situation with unknown variance. Positive values of $\theta$ imply superiority of the treatment $T$. Suppose the block size at each stage to be $\{2B_i, i = 1, 2, \ldots\}$, where $B_i$ for each arm is pre-fixed. Let $\overline{X}_i = \overline{X}_{T_i} - \overline{X}_{C_i}$ denote the mean difference for the $i$th block of data. $\overline{X}_i \sim N(\theta, \sigma^2/B_i)$.

Let $S_i$ define the standardized statistic based on the $i$th block of data and $U_j$ define the standardized statistic for the cumulated data up to the $j$th analysis; then

$$S_i = B_i^{1/2}\overline{X}_i/\sigma \sim N(B_i^{1/2}\theta/\sigma,\ 1), i = 1, 2, \ldots, m,$$

$$U_j = \sum_{i=1}^{j} B_i^{1/2} S_i / n_j^{1/2} \sim N(n_j^{1/2}\theta/\sigma,\ 1),$$

$$n_j = \sum_{i=1}^{j} B_i,\ \textit{for } j = 1, \ldots, m.$$

The one-sided null hypothesis, $H_0 : \theta \leq 0$, is tested against the alternative hypothesis $H_1 : \theta > 0$.

The trial is reviewed after observing every $2B_i$ subjects at the $i$th interim analysis. If the cumulated information shows sufficient evidence that the new treatment is ineffective or inferior to the standard one, the trial should be terminated early. At each interim analysis, the futility boundary is constructed from the confidence limit of $\theta$ specified at significance level $\alpha_0$ and the expected mean difference $\delta$. If $U_j$ is below the futility boundary, the trial is terminated to accept $H_0$ at the next step. Otherwise, a weight based on the conditional power is calculated using the cumulated data before the current step. The procedure is iterated until the total squared weight reaches one at step $m$, that is, $\sum_{i=1}^{m-1} w_i^2 < 1$ and $\sum_{i=1}^{m} w_i^2 \geq 1$. Specifically, after observing the $j$th block of data,

(i) If $U_j \leq \frac{n_j^{1/2}\delta}{\sigma} - z_{\alpha_0}$, let $m = j + 1$ and accept $H_0$. If $U_j > \frac{n_j^{1/2}\delta}{\sigma} - z_{\alpha_0}$ and $\sum_{i=1}^{j+1} w_i^2 \geq 1$, let $m = j + 1$ and the trial terminates at the $j + 1$th block.

(ii) If $U_j > \frac{n_j^{1/2}\delta}{\sigma} - z_{\alpha_0}$ and $\sum_{i=1}^{j+1} w_i^2 < 1$, observe the next block of data and repeat (i) and (ii).

The weight functions are constructed using conditional power except at the first step. Specifically, $w_1 = (B_1/N_1)^{1/2}$, where $N_1 = (z_\alpha + z_\beta)^2\sigma^2/\delta^2$, and $\alpha$ and $\beta$ are the specified Type I and II error rates, respectively. Note that $\alpha_0$ is the significance level used for futility monitoring that can be different from $\alpha$. For $j \geq 2$,

$$w_j = \left\{ B_j(1 - \sum_{i=1}^{j-1} w_i^2)/N_j \right\}^{1/2}, \text{where}$$

$$N_j = \left\{ (z_\alpha - \sum_{i=1}^{j-1} w_i S_i)/(1 - \sum_{i=1}^{j-1} w_i^2)^{1/2} + z_\beta \right\}^{1/2} \cdot \sigma^2 / \hat{\theta}_{j-1}^2.$$

$\hat{\theta}_j = U_j\sigma/\sqrt{n_j}$ is the naive estimator of $\theta$ using the cumulated data up to the $j$th block, and $N_j$ is solved from an inequality aimed at achieving conditional power of $1 - \beta$. The final test statistic is constructed as a weighted average of the sufficient statistic of each block:

$$T_m = \sum_{i=1}^{m} w_i S_i,$$

which follows the standard normal distribution when $\theta = 0$. If $T_m > Z_\alpha$, $H_0$ is rejected; otherwise $H_0$ is accepted at the final inference, where $w_m^2$ is redefined as $1 - \sum_{i=1}^{m-1} w_i^2$.

## The Distribution of the Stopping Time

A trial is terminated at the $m$th analysis if and only if one of the following two events occurs: The cumulated data cross the futility boundary, or the weight is used up at the $m$th analysis, while the

summation of the squared weights is less than 1 and the futility boundary is not crossed for any $j < m$; mathematically, that is

$$\{m = k\} = \left\{U_j > n_j^{1/2}\delta/\sigma - z_{\alpha_0},\ j \le k-2, \sum_{i=1}^{k-1} w_i^2 < 1\right\} \bigcap \left[\{U_{k-1} \le n_{k-1}^{1/2}\delta/\sigma - z_{\alpha_0}\} \bigcup \{U_{k-1} > n_{k-1}^{1/2}\delta/\sigma - z_{\alpha_0}, \sum_{i=1}^{k} w_i^2 \ge 1\}\right]. \quad (1)$$

Decomposing Equation 1 to a sequence of continuation regions at interim analyses facilitates the computation of this probability. When $m = k$, let

$$\begin{aligned} E_{k,1} &= \{U_1 > n_1^{1/2}\delta/\sigma - z_{\alpha_0},\ {w_1}^2 + {w_2}^2 < 1\},\\ E_{k,j} &= \{U_j > n_j^{1/2}\delta/\sigma - z_{\alpha_0},\ {w_1}^2 + \cdots + {w_{j+1}}^2 < 1\}, \text{for } j = 2, \ldots, k-2,\\ E_{k,k-1} &= \{U_{k-1} \le n_{k-1}^{1/2}\delta/\sigma - z_{\alpha_0}\} \bigcup \{U_{k-1} > n_{k-1}^{1/2}\delta/\sigma - z_{\alpha_0}, \sum_{i=1}^{k} w_i^2 \ge 1\}, \end{aligned}$$

where event $E_{k,j}$ only depends on the data from the first $j$ blocks. It is clear that Equation 1 is equivalent to the intersection of $E_{k,j}$ for $j = 1, \ldots, k-1$. The probability mass function for the stopping time $m$ can be expressed as

$$\begin{aligned} P_\theta(m = k) &= P_\theta\left(\bigcap_{j=1}^{k-1} E_{k,j}\right)\\ &= P_\theta(E_{k,1}) \prod_{j=2}^{k-1} P_\theta(E_{k,j} | E_{k,1}, \ldots, E_{k,j-1}). \end{aligned} \quad (2)$$

The expression on the right-hand side of Equaton 2 makes the computation tangible because each conditional probability can be evaluated recursively. To calculate the probability and each of the conditional probabilities on the right-hand side of Equation 2, define

$$b_j = \frac{n_j(\delta - \theta)}{B_j^{1/2}\sigma} - z_{\alpha_0}\left(\frac{n_j}{B_j}\right)^{1/2},\quad Z_j = S_j - \frac{B_j^{1/2}\theta}{\sigma}, \text{and}$$

$$Y_{j+1} = \sum_{i=1}^{j} \left(\frac{B_i}{B_{j+1}}\right)^{1/2} Z_i. \quad (3)$$

It can be verified that $Z_1, Z_2, \ldots$, are independent standard normal variables. The advantage for the transformation is to simplify the derivation of the joint distribution $\{U_1, \ldots, U_k\}$ to the derivation of the distribution of $\{Z_1, \ldots, Z_k\}$. The region where the weight is not yet used up at the $j$th analysis is

$$D_j = \left\{(z_1, \ldots, z_j) :\quad \sum_{i=1}^{j+1} w_i^2(s_1, \ldots, s_{i-1}) < 1,\ s_i = z_i + \frac{B_i^{1/2}\theta}{\sigma}\right\} \quad (4)$$

for $j = 1, 2, 3, \ldots, m-1$. Continuation region $E_{k,j}$ ($j < k-1$) can be expressed by an intersection of the two events in terms of $Z_j$ and $D_j$ as follows:

$$\begin{aligned} &E_{k,1} = \{Z_1 > b_1\} \bigcap D_1,\\ &E_{k,j} = \{Z_j > -Y_j + b_j\} \bigcap D_j,\ \textit{for } j = 2, \ldots, k-2. \end{aligned}$$

The termination region consists of the following two exclusive events: The futility boundary is crossed, or the weight is used up,

$$E_{k,k-1} = \{Z_{k-1} \le -Y_{k-1} + b_{k-1}\} \bigcup \left[\{Z_{k-1} > -Y_{k-1} + b_{k-1}\} \bigcap \overline{D}_{k-1}\right],$$

where $\overline{D}_j$ is the complement of $D_j$. Given $\{Z_1, \ldots, Z_{j-1}\}$, the computation of continuation region $E_{k,j}$ depends only on $Z_j$ for $j < k$. The probability mass function of $m$ in Equation 2 can then be formulated by the following integral:

$$P_\theta(m = k) = \int_{E_{k,1}} \phi(z_1) \int_{E_{k,2}} \phi(z_2) \ldots \int_{E_{k,k-1}} \phi(z_{k-1}) dz_{k-1} \ldots dz_1,$$

where $\phi(\cdot)$ is the probability density function of the standard normal distribution.

## Evaluation of Operating Characteristics

Fundamental quantities for evaluating a group sequential design include the probability of exiting and rejecting (or accepting) the null hypothesis at a given interim analysis, and the overall probability to reject or accept the null hypothesis with the given design parameters and $\theta$.

To compute the probability of exiting and rejecting $H_0$ at the $k$th analysis, express the event to terminate the trial and reject $H_0$ at $m = k$ as

$$R_k = \left\{U_j > n_j^{1/2}\delta/\sigma - z_{\alpha_0},\ 1 \le j \le k-1, T_k > z_\alpha\right\} \bigcap \overline{D}_{k-1} \bigcap_{i=1}^{k-2} D_i,$$

where $D_i$ is defined in Equation 4. Moreover, $R_k$ is equal to the intersection of the following sequential events:

$$\begin{aligned}
C_{k,1} &= \{U_1 > n_1^{1/2}\delta/\sigma - z_{\alpha_0}, w_1^2 + w_2^2 < 1\} \\
&= \{Z_1 > b_1\} \bigcap D_1, \\
C_{k,j} &= \{U_j > n_j^{1/2}\delta/\sigma - z_{\alpha_0},\ w_1^2 + \cdots + \\
&\qquad w_{j+1}{}^2 < 1\} = \{Z_j > b_j - Y_j\} \bigcap D_j, \\
C_{k,k-1} &= \{U_{k-1} > n_{k-1}^{1/2}\delta/\sigma - z_{\alpha_0}, w_1{}^2 + \cdots + \\
&\qquad w_k{}^2 \ge 1\} = \{Z_{k-1} > b_{k-1} - Y_{k-1}\} \bigcap \overline{D}_{k-1}, \\
C_{k,k} &= \{T_k > z_\alpha\} \\
&= \left\{Z_k > \frac{z_\alpha}{w_k} - \sum_{i=1}^{k-1} \frac{w_i}{w_k} Z_i - \sum_{i=1}^{k} \frac{w_i B_i^{1/2}\theta}{w_k \sigma}\right\}
\end{aligned}$$

for $j = 2, \ldots, k-2$, where $C_{k,j}$ is the continuation region for $j < k-1$ and $C_{k,k}$ is the rejection region at the $k$th analysis. $C_{k,k-1}$ shows that the weight is used up to the $k$th analysis. Event $C_{k,j}$ depends only on data from the first $j$ blocks for $j \le k$.

The probability of a trial being terminated at the $k$th analysis and rejecting the null hypothesis can be computed by the product of the following conditional probabilities:

$$\begin{aligned}
P_\theta(R_k) = P_\theta\left(\bigcap_{j=1}^{k} C_{k,j}\right) = P_\theta(C_{k,1}) \prod_{j=2}^{k} \\
P_\theta(C_{k,j} | C_{k,1}, \ldots, C_{k,j-1}).
\end{aligned} \tag{5}$$

Because $Z_1, Z_2, \ldots, Z_k$ are independent identically distributed standard normal variables and each $C_{k,j}$ is uniquely determined by $Z_j$ conditional on $Z_1, \ldots, Z_{j-1}$, Equation 5 can be written as a $(k-1)$-fold multiple integral:

$$\begin{aligned}
P_\theta(R_k) = \int_{C_{k,1}} \phi(z_1) \cdots \int_{C_{k,k-2}} \phi(z_{k-2}) \cdot \\
\int_{C_{k,k-1}} \phi(z_{k-1}) \Phi\left(\sum_{i=1}^{k-1} \frac{w_i}{w_k} Z_i + \sum_{i=1}^{k} \frac{w_i B_i^{1/2}\theta}{w_k \sigma} - \frac{z_\alpha}{w_k}\right) \\
dz_{k-1} \ldots dz_1,
\end{aligned}$$

where $\Phi(\cdot)$ is a cumulative distribution function of a standard normal variable.

The overall rejection region, $R$, can be decomposed by the partition $\{m = k\}$, for $k = 2,3, \ldots$. That is, $R = \bigcup_{k=2}^{\infty} R_k$, where $R_k = R \bigcap \{m = k\}$. Consequently, the overall probability of rejecting the null hypothesis can be calculated by the following summation:

$$P_\theta(R) = \sum_{k=2}^{\infty} P_\theta(R_k).$$

If a trial is terminated to accept $H_0$ at a given interim analysis, $m = k$, the trial is terminated either by spending all the weight and $T_k \le z_\alpha$ at the $k$th analysis or by crossing the futility boundary at the $k-1$th analysis.

$$\begin{aligned}
A_k &= \{U_j > n_j^{1/2}\delta/\sigma - z_{\alpha_0}, 1 \le j \le k-1, \\
&\qquad T_k \le z_\alpha,\ m = k\} \\
&\quad \bigcup \{U_j > n_j^{1/2}\delta/\sigma - z_{\alpha_0},\ 1 \le j \le k-2, \\
&\qquad U_{k-1} \le n_{k-1}^{1/2}\delta/\sigma - z_{\alpha_0}\} \\
&= \left\{\left(\bigcap_{j=1}^{k-1} C_{k,j}\right) \bigcap \overline{C}_{k,k}\right\} \\
&\qquad \bigcup \left\{\left(\bigcap_{j=1}^{k-2} C_{k,j}\right) \bigcap (Z_{k-1} \le b_{k-1} - Y_{k-1})\right\}.
\end{aligned}$$

The probability of accepting $H_0$ at the $k$th analysis can be computed by the following multiple integral:

$$P_\theta(A_k) = \int_{C_{k,1}} \phi(z_1) \cdots \int_{C_{k,k-2}} \phi(z_{k-2}) \cdot \int_{C_{k,k-1}} \phi(z_{k-1}) \Phi\left(\frac{z_\alpha}{w_k} - \sum_{i=1}^{k-1} \frac{w_i}{w_k} Z_i - \sum_{i=1}^{k} \frac{w_i B_i^{1/2} \theta}{w_k \sigma}\right) dz_{k-1} \cdots dz_1,$$

$$+ \int_{C_{k,1}} \phi(z_1) \cdots \int_{C_{k,k-2}} \phi(z_{k-2}) \cdot \Phi(b_{k-1} - Y_{k-1}) dz_{k-2} \ldots dz_1.$$

Subsequently, the probability that a trial terminates at the $k$th analysis is $P_\theta(A_k) + P_\theta(R_k)$. For any given $\theta$, the overall probability of accepting the null hypothesis can be obtained by

$$P_\theta(A) = \sum_{k=2}^{\infty} P_\theta(A_k).$$

## Applications and Numerical Results

The properties derived in the last two sections can be used to design a self-designing sequential trial and to obtain directly the error probabilities for a given set of design parameters. The values of the design parameters $(\alpha, \beta, \alpha_0, \delta)$ determine the distribution of the stopping time and the associated decision rules for any given $\theta$. For the fixed sample design, the actual Type I error probability would be $\alpha$ if $\theta = 0$, and the actual Type II error probability would be $\beta$ if $\theta = \delta$, whereas the actual error probabilities might be different if $\theta \neq \delta$. Similarly, in a self-designing trial, the actual overall Type I or II error rate is often different from the specified $\alpha$ or $\beta$, given different values of $\theta$. Using the derived formulas, the actual error rates can be estimated for any given value of $\theta$ and the probability of rejecting or accepting the null hypothesis at a given interim analysis can be obtained.

Table 1 presents the numerical results given $\alpha = 0.05$, $\beta = 0.1$, $\alpha_0 = 0.05$, and $\delta = 0.4$, 0.5, or 0.6, with the true mean $\theta = 0$, 0.5, or 0.7, and prefixed constant block sizes, $B_1 = N_1/2$ and $B_i = 10$ or 12 for $i \geq 2$.

There is a rapid decay of the probability mass $P_\theta(m = k)$ when $k > 3$ in general. As presented in Table 1 for $P_\theta(m \leq 3)$, 85% to 97% of trials were terminated with three or fewer interim analyses either under $H_0$ or $H_1$ for the given block sizes. When $P_\theta(m > 3)$ is small, the overall probability of rejecting $H_0$, $P_\theta(R)$ can be approximated by the following conditional probability:

$$P_\theta(R|m \leq 3) = \frac{P_\theta(R_2) + P_\theta(R_3)}{P_\theta(m = 2) + P_\theta(m = 3)}, \quad (6)$$

**Table 1** Probability of Stopping Time, Conditional Probability of Rejecting $H_0$ for $m \leq 3$, and the Simulated Probability of Rejection When $\alpha = 0.05$, $\beta = 0.1$

| | | *B=10* | | | *B=12* | | |
|---|---|---|---|---|---|---|---|
| $\theta$ | $\delta$ | $P_\theta(m \leq 3)$ | *C.P.* | *S.P.* | $P_\theta(m \leq 3)$ | *C.P.* | *S.P.* |
| 0.0 | 0.4 | 0.845 | 0.022 | 0.025 | 0.872 | 0.024 | 0.023 |
| | 0.5 | 0.896 | 0.025 | 0.028 | 0.927 | 0.027 | 0.027 |
| | 0.6 | 0.945 | 0.027 | 0.028 | 0.972 | 0.029 | 0.028 |
| | 0.7 | 0.973 | 0.029 | 0.026 | 0.991 | 0.030 | 0.029 |
| 0.5 | 0.4 | 0.870 | 0.952 | 0.954 | 0.906 | 0.955 | 0.955 |
| | 0.5 | 0.851 | 0.864 | 0.876 | 0.902 | 0.876 | 0.875 |
| | 0.6 | 0.883 | 0.765 | 0.759 | 0.934 | 0.777 | 0.779 |
| | 0.7 | 0.929 | 0.656 | 0.657 | 0.971 | 0.673 | 0.661 |
| 0.7 | 0.5 | 0.947 | 0.984 | 0.985 | 0.961 | 0.987 | 0.989 |
| | 0.6 | 0.938 | 0.953 | 0.955 | 0.959 | 0.958 | 0.962 |
| | 0.7 | 0.948 | 0.892 | 0.896 | 0.964 | 0.905 | 0.911 |
| | 0.8 | 0.964 | 0.831 | 0.838 | 0.967 | 0.842 | 0.836 |

*Note:* $B_1 = 0.5N_1$, $B_i = B$ for $i \geq 2$, C.P. $= P_\theta(R|m \leq 3)$, and S.P. is the simulated probability of rejection based on 10,000 repetitions.

**Table 2** The Value of $\alpha_0$ Determined to Achieve Power 0.9 With $\alpha = 0.05$

| | | *B = 8* | | | *B = 10* | | | *B = 12* | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\theta$ | $\delta$ | $\alpha_0$ | *C.P.* | *S.P.* | $\alpha_0$ | *C.P.* | *S.P.* | $\alpha_0$ | *C.P.* | *S.P.* |
| 0.4 | 0.4 | 0.015 | 0.906 | 0.901 | 0.015 | 0.908 | 0.909 | 0.020 | 0.901 | 0.911 |
| | 0.5 | 0.0015 | 0.903 | 0.861 | 0.0015 | 0.900 | 0.869 | 0.0015 | 0.903 | 0.879 |
| 0.5 | 0.5 | 0.025 | 0.902 | 0.902 | 0.025 | 0.906 | 0.908 | 0.030 | 0.903 | 0.903 |
| | 0.6 | 0.003 | 0.906 | 0.879 | 0.004 | 0.902 | 0.876 | 0.007 | 0.906 | 0.866 |
| 0.6 | 0.6 | 0.035 | 0.900 | 0.901 | 0.040 | 0.901 | 0.906 | 0.045 | 0.902 | 0.907 |
| | 0.7 | 0.006 | 0.903 | 0.886 | 0.005 | 0.911 | 0.897 | 0.005 | 0.917 | 0.906 |

because $P_\theta(R)=P_\theta(R|m\leq 3)P_\theta(m\leq 3)+P_\theta(R|m>3)P_\theta(m>3)$.

The performance of this approximation is assessed by comparing the conditional probability (Equation 6) calculated from the numerical integration, denoted by C.P. in Table 1, with the empirical overall probability of rejecting the null hypothesis, denoted by S.P., where S.P. is estimated using the Monte Carlo method with 10,000 repetitions. The quantities under "S.P." in Table 1 are the empirical estimators of Type I error rates (under $\theta = 0$) or power (under $\theta > 0$). The associated estimator under "C.P." is the conditional probability directly calculated from the numerical integration. As expected, the conditional probability is a reasonable approximation to $P_\theta(R)$ with a moderate block size ($B \geq 10$) under either $H_0$ or $H_1$.

It is clear that the choice of $\alpha_0$ determines the futility boundary, which affects the operating characteristics of the design. The largest possible $\alpha_0$ is searched to achieve a power $(1 - \beta)$ with a specified Type I error rate $\alpha$ by setting $P(R|m \leq 3) \geq 1 - \beta$.

Presented in Table 2 with $\alpha = 0.05$ and $\theta = \delta$, $\alpha_0$ varies from 0.0015 to 0.045 to attain the specified power, 0.9, depending on the values of $\delta$ and $\alpha$, and the block size. If the true mean $\theta$ is smaller than the expected mean $\delta$, a trial with design parameter $\alpha_0$, which is determined under the assumption of $\theta = \delta$, might not achieve the desired power. Instead, a much smaller $\alpha_0$ for the futility boundary is needed to achieve the specified power.

The evaluations of the operation characteristics can help investigators to assess the feasibility of a flexible adaptive design easily under some scenarios of interest. The formulation also allows clinical trialists to select the design parameters so as to construct a sequential design to satisfy regulatory requirements. In particular, the methodology described here can help one to better achieve the targeted Type I and II errors.

*Yi Cheng and Yu Shen*

***See also*** Adaptive Designs in Clinical Trials; Block Design; Clinical Trial; Error Rates; Group-Sequential Designs in Clinical Trials; Power; Sample Size Planning

**Further Readings**

Cheng, Y., & Shen, Y. (2004). Estimation of a parameter and its exact confidence interval following sequential sample size reestimation trials. *Biometrics, 60,* 910–918.

Jennison, C., & Turnbull, B. W. (2006). Adaptive and nonadaptive group sequential tests. *Biometrika, 93,* 1–21.

Lehmacher, W., & Wassmer, G. (1999). Adaptive sample size calculations in group sequential trials. *Biometrics, 55,* 1286–1290.

Müller, H. H., & Schäfer, H. (2001). Adaptive group sequential designs for clinical trials: Combining the advantages of adaptive and of classical group sequential approaches. *Biometrics, 57,* 1034–1041.

Shen, Y., & Cai, J. (2003). Sample size reestimation for clinical trials with censored survival data. *Journal of the American Statistical Association, 98,* 418–426.

Shen, Y., & Fisher, L. (1999). Statistical inference for self-designing clinical trials with a one-sided hypothesis. *Biometrics, 55,* 190–197.

# Sequential Design

Experimental studies in the social and behavioral sciences typically follow a *fixed* experimental

design where the sample size and composition (e.g., experimental group allocation) is determined prior to conducting the experiment. In contrast, *sequential* experimental designs treat the sample size as a random variable by allowing sequential interim analyses and decision making based on cumulative data and previous design decisions while maintaining appropriate control over experiment-wise errors in decision making (i.e., Type I [$\alpha$] and Type II [$\beta$] error rates). Also referred to as *adaptive* or *flexible* designs, current design decisions in sequential designs are sequentially selected according to previous design points. Sequential designs rely on the principal of *stochastic curtailment* to stop the experiment if the given data at an interim analysis are likely to predict the eventual outcome with a high probability. In contrast to a fixed design where the size and composition of the sample is determined prior to conducting the experiment, the number of observations or participants is not predetermined in a sequential design. Thus, the size and composition of the sample is considered random because of decision dependence on previous observations. However, a finite upper limit is often set in practice.

Sequential designs allow for the early termination of experiments if cumulative evidence suggests a clear effect or lack thereof. The capacity for sequential designs to terminate early can be beneficial from an ethical perspective by preventing unnecessary exposure to unsafe experimental conditions in terms of both length of exposure and the number of participants exposed, as well as unnecessarily withholding administration when the experimental condition is clearly beneficial. Popular in medical trials and drug studies, the early termination of prevention or intervention experiments might lead to minimized exposure to potentially harmful or ineffective treatments. Sequential designs can also be beneficial from a logistical perspective as they can save both time and resources. More generally, sequential designs have the potential to lead to financial savings because of reduced sample sizes. Under the null hypothesis of no effect, sequential designs might be stopped for lack of effectiveness at a total sample size smaller than would be the case with a fixed design. Under the alternative hypothesis of efficacious experimental manipulation, a similar savings is observed in the total sample size required, with the sample size savings typically reported as greater under the alternative hypothesis than under the null hypothesis. Actual sample reductions vary by the type of sequential design and certain design decisions but are generally reported to be as large as 10% under the null hypothesis and as large as 50% under the alternative hypothesis. In research contexts where the participant stream might be small or inaccessible, the potential for such substantial sample size savings when implementing a sequential experimental design might be quite beneficial.

This entry discusses the background and design characteristics, along with the benefits and limitations of sequential design.

## Background

The origins of sequential designs can be traced back to the development of a double sampling inspection procedure by Harold F. Dodge and Harry G. Romig in 1929 for the purpose of industrial quality control. Prasanta Chandra Mahalanobis's 1938 census of Bengalese jute area is also considered an important precursor. In 1943, Abraham Wald, with collaboration from members of the Statistical Research Group at Columbia University, including Milton Friedman and W. Allen Wallis, developed the sequential probability ratio test for military armament testing. The development of the sequential probability ratio test also launched the complementary field of sequential analysis. Sequential analyses are statistical hypothesis testing procedures that allow a statistical test to be calculated at any stage of the experiment prior to completion, and then it provides a three-alternative rule for inferential decision making: to fail to reject the null hypothesis, to reject the null hypothesis, or to continue the experiment. In contrast, statistical evaluation of a fixed design experiment only occurs at the completion of the experiment. Peter Armitage's 1960 book on sequential medical trials effectively introduced the sequential design of randomized clinical trials (RCT), and in this research context, sequential designs have observed the greatest growth and development. More recently, psychometricians have developed computerized adaptive testing procedures for educational and psychological testing based on the principles of sequential

design of experiments. Psychometric applications of sequential design became widespread beginning in the early 1980s, but its roots can be attributed to Alfred Binet in 1905 with the start of adaptive individualized intelligence testing.

## Design Characteristics

All sequential designs provide the opportunity for at least one interim analysis at a prespecified interim stage prior to formal completion of the experiment, which allows early stopping for either positive or negative results through progress monitoring. Prior to the start of the experiment, the statistical details are determined, including the number of interim stages and, relatedly, the sample size at each stage. Based on the desired nominal $\alpha$ and $\beta$ levels, critical values, which are also called *boundary values,* are computed for each interim stage. At each interim stage, all available data are analyzed including data from that stage plus all previous stages. The appropriate test statistic and the Fisher information level (the inverse of the squared standard error) are computed at each stage. The test statistic is then compared with critical boundary values chosen a priori to maintain appropriate nominal experiment-wise Type I and Type II error rates given the occurrence of multiple statistical tests at interim stages. If the test statistic falls within a decision region, the experiment stops. Otherwise, the experiment continues to the next stage or until the maximum sample size is reached.

### *Boundary Values*

At any interim stage, the experiment can either be stopped for ethical or efficacy reasons, or continued as a result of insufficient statistical evidence of a significant effect in either direction. Boundary values similar to conventional critical values are set up for each interim stage to determine whether sufficient accumulated evidence is available to conclude experimental efficacy (or futility) or the presence of a harmful effect. Boundary values are derived to maintain experiment-wise Type I and Type II error rates at specified nominal levels. Up to four boundary values are derived a priori depending on the number of tails in the statistical hypothesis test. A one-tailed test has up to two boundary values, while a two-tailed hypothesis has up to four boundary values. Two boundary values define the region of rejection where the experiment stops and the null hypothesis of no difference is rejected. Assuming that the experimental group experiences an increase in outcomes as a result of a particular treatment, if the test statistic exceeds the upper critical value, the null hypothesis is rejected for efficacy. However, if the test statistic exceeds the lower critical value, the null hypothesis is rejected for a harmful effect. Two additional boundary values can be specified to define the acceptance region where the experiment stops and the null hypothesis is not rejected. If the test statistic falls in between the rejection and acceptance regions, the experiment continues to the next interim stage. At the last stage, the boundaries of the rejection and acceptance regions are identical, the experiment stops, and the decision is made either to reject or fail to reject the null hypothesis. If the experimental protocol only allows early stopping to reject, there are no acceptance boundaries at early stages, and with early stopping only to fail to reject the null hypothesis, there are no interim rejection boundaries.

### *Boundary Methods*

Traditional fixed design methods for determining critical or boundary values cannot be applied in sequential designs because of inflation in the Type I error rate. For instance, a fixed design testing a two-tailed hypothesis where a $z$ test only occurs at the completion of the experiment would use a pair of critical values defining the boundary of the region of rejection of the null hypothesis, $\pm 1.96$, to control the experiment-wise Type I error rate at the nominal $\alpha = .05$ level. A two-stage sequential design that was required to be carried out to completion rather than stopped early after one interim analysis, resulting in the same sample size as the fixed design, would have a nominal Type I error rate of $\alpha = .083$ using the same critical values because of the experiment consisting of two dependent statistical tests. Consequently, multiple methods have been developed to control the Type I error rate adequately in sequential experiments. These methods are of three general types: fixed boundary shape methods, Whitehead methods, and error spending methods.

The *fixed boundary shape methods* include several popular approaches, including the Pocock and O'Brien-Fleming methods. Use of the O'Brien-Fleming method requires overwhelming evidence to reject the null hypothesis at early stages by implementing initially conservative, but successively decreasing, boundary values. For instance, a five-stage design would use critical values of 4.56, 3.23, 2.63, 2.28, and 2.04, where the final critical value is then close to the fixed design critical $z$ value of 1.96. In contrast, Pocock's method, when used with equally spaced information levels, derives a constant boundary value, but the nominal alpha level is smaller than the desired $\alpha = .05$ level, which makes the overall design more statistically conservative. For instance, a constant boundary value at all five stages would be 2.41 rather than 1.96, but the nominal alpha level at the final stage ($\alpha = .032$) would be smaller than the overall alpha-level of the design ($\alpha = .050$). Such a discrepancy might result in a failure to reject the null hypothesis despite the final $p$ value falling below the nominal experiment-wise level (i.e., if the final $p$ value is between .032 and .050).

*Whitehead methods* were generalized from designs requiring continuous monitoring (i.e., after each participant completes the study) to designs with discrete monitoring (i.e., after groups of participants complete the study). They result in triangular or straight-line boundaries, require less computation that the fixed boundary shape methods, and maintain Type I error rates and statistical power that is close to, but does differ slightly from, nominal values.

The third class of boundary methods, the *error spending methods,* requires the most computation (relative to Whitehead and fixed boundary shape methods) when a large number of interim stages are used. Error spending methods use an error spending function to specify an amount of the prespecified nominal alpha and beta rate error to be spent at each stage with boundary values derived from this amount.

### *Types of Sequential Designs*

Sequential experimental designs are generally of three types that differ based on sample recruitment and decision-making criteria.

*Fully sequential designs* use continuous monitoring and are updated after every observation or after every participant completes the study; thus, the critical boundary values change as the experiment progresses because of the addition of data and a resulting reduction of variance. Fully sequential designs require that the outcome of the experiment is knowable in a relatively short period of time relative to the length of time required for recruitment or follow-up. In such designs, long periods of "no effect" or continuation might be falsely interpreted as equivalence of experimental conditions. Although Abraham Wald's development of the sequential probability ratio test was the origin of this approach, despite its historical longevity, it has not been widely applied. However, one area where fully sequential designs have been widely applied is in the context of computerized adaptive testing (CAT) and computerized classification testing (CCT), where question selection depends on the response to the previous question. In these contexts, each question that a participant responds to is considered an observation. The test (i.e., the experiment) concludes when a prespecified degree of precision or consistency of responses is attained rather than a statistically significant difference between two experimental conditions.

*Group sequential designs* have been found to be more practical in most settings than fully sequential designs. In some circumstances, an evaluation after each participant completes the study might not be feasible. Also, choosing to stop an experiment early after just a few participants because of strong evidence might not be persuasive either way the evidence falls. Group sequential designs can be considered analogous to fully sequential designs with the distinction that boundary values are computed for a predetermined number of equally spaced stages based on a prespecified number of participants (i.e., a group) rather than after each participant. Likewise, fully sequential designs can be viewed as group sequential designs where the size of the participant group is $n = 1$. Four to five interim stages or data inspections are typically recommended, as a larger number of interim analyses do not lead to extra benefits in terms of reduced overall sample sizes.

Whereas fully sequential and group sequential designs are characterized by prespecified, predetermined, and constant group sizes (with $n = 1$ as

a special case), *flexible sequential designs* allow flexible modification during the experiment through an alpha spending function that allocates the amount of the nominal experiment-wise error to be spent at each interim while maintaining the overall nominal Type I error rate at the $\alpha = .05$ level. To protect against potential researcher abuse, the alpha spending function is specified during the study design phase and cannot be changed midexperiment. Flexible sequential designs allow periodic interim evaluation as do other types of sequential designs, but flexible designs also might allow the interim evaluations to become more frequent as the decision point becomes closer. Thus, flexible sequential designs can be viewed as a compromise between fully sequential and group sequential designs where each is considered a special case of a more general flexible design. The flexible design is equivalent to the fully sequential design under continual monitoring and equivalent to the group sequential design when the flexible nature of the interim analyses is not used.

## Benefits and Limitations

The primary engine behind the principles of sequentially designed experiments is making use of existing information at interim stages of the experiment to inform future design decisions, especially in regard to participant recruitment. In the case where the researcher has a limited amount of theoretical or empirical knowledge that would otherwise prevent well-informed decision making in terms of the necessary sample size, sequential designs can be quite valuable. The advantage of a sequential experimental design as compared with a fixed design is primarily tied to the size of the experimental effect in the population. Traditional designs are planned to have sufficient power to detect an effect size specified in advance by the researcher. To the extent that this effect size is incorrect in the population, the researcher has inadvertently overpowered or underpowered the study; both situations are problematic. This brings to light a specific advantage of sequential designs over fixed experimental designs. A sequential design allows early stopping in the event the effect size is larger than previously estimated, as well as subsequent data collection in the event the effect size is smaller than expected. If experimental differences are found to be larger than expected, the probability of early study termination is higher and savings might be obtained in terms of decreased resource allocation (smaller samples, fewer materials, etc.) regardless of the general type of sequential design employed. If the effect of interest is actually smaller in the population than expected, then fully sequential and group sequential designs might result in a decrease in power because of the additional analyses. This result can be ameliorated to an extent by employing a flexible sequential design. For instance, a particular experiment is designed so that a given effect size has a prespecified probability of being detected by the end of the study (power) while controlling the probability of making a Type I error at the nominal $\alpha = .05$ level. However, at an interim stage the effect size might be smaller than expected, which raises the concern that there might be a lack of power for the overall experiment. If the overall design is flexible, the sample size might be increased and the boundary values adjusted to preserve the Type I error rate. This approach would not be as statistically efficient as a group sequential design that does not adapt, but it would protect against insufficient statistical power to reject a null hypothesis that is truly false. Some might argue that adopting a flexible design from the onset of a study is a reasonable approach when the magnitude of the true treatment effect is not clearly known, if there is a discrepancy between a clinically meaningful effect and an observable effect, or if it is not clear that the potential effectiveness warrants using what are otherwise limited resources. In this circumstance, a flexible design would allow accumulation of promising evidence to support better-informed expansion or continuation of the experiment.

Sequential designs are particularly applicable to medical and pharmaceutical trials, which are generally referred to as randomized clinical trials (RCTs). An RCT is a type of randomized experiment in which the effectiveness of a new drug, intervention, or other medical procedure is evaluated by comparing a treatment group with a control group. RCTs are conducted according to a plan, or *protocol,* which details the RCT's objectives, data collection procedures, and data analysis framework. RCTs typically require a sequentially recruited participant stream and lend themselves

well to sequential designs. Sequential designs are particularly appropriate for medical RCTs as well because of the need for trial monitoring to minimize exposure to unacceptable toxicity or potential harm of new interventions or to minimize continuation after the benefit or risk is clearly apparent. Particularly when the clinical outcome is considered irreversible, fixed designs are not acceptable; thus, monitored sequential designs are needed.

Flexible sequential designs are particularly useful in Phase II clinical trials. Whereas a Phase I trial is used to determine whether a new drug or treatment is safe for humans and to estimate an initial effect size, a Phase II trial is conducted on a larger sample of participants to determine how well the drug works while continuing to monitor participant safety. Phase I effect size estimates are sometimes found to be too optimistic, so a Phase II sample size based on an inflated effect size might be smaller than required. A flexible sequential design would allow for interim sample size augmentation to accommodate the smaller-than-expected effect size.

Finally, sequential designs are not without their limitations. In addition to the increased design complexity and computational burdens in determining boundary values and controlling the experiment-wise error rate, the ability to terminate the experiment early (a major benefit of sequential designs) also provides a threat to the validity of the study. If an experiment stops early for efficacy, futility, or participant safety, a smaller sample size can lead to a distrust of the findings because most statistical procedures rely on asymptotic principles that require larger sample sizes. Thus, the decision to exit early is more complex than just a statistical criterion. Therefore, the sequential design plan must be specified so that the resulting analysis is both reliable and persuasive. This includes consistency across both primary and secondary outcomes, risk groups, and so on, to be persuasive after an early termination.

*James A. Bovaird and Kevin A. Kupzyk*

***See also*** Clinical Trial; Computerized Adaptive Testing; Error Rates; Experimental Design; Group-Sequential Designs in Clinical Trials; Sample Size Planning; Sequential Analysis; "Sequential Tests of Statistical Hypotheses"

**Further Readings**

Armitage P. (1975). *Sequential medical trials* (2nd ed.). New York: Wiley.

Chernoff, H. (1959). Sequential design of experiments. *The Annals of Mathematical Statistics, 30,* 755–770.

DeMets, D. L. (1998). Sequential designs in clinical trials. *Cardiac Electrophysiology Review, 2,* 57–60.

DeMets, D. L., & Lan, K. K. (1994). Interim analysis: The alpha spending function approach. *Statistics in Medicine, 13,* 1341–1352.

Muller, H. H., & Shafer, H. (2001). Adaptive group sequential designs for clinical trials: Combining the advantages of adaptive and of classical group sequential approaches. *Biometrics, 57,* 886–891.

O'Brien, P. C., & Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics, 35,* 549–556.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika, 64,* 191–200.

Wainer, H. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics, 16,* 117–186.

Whitehead, J., & Stratton, I. (1983). Group sequential clinical trials with triangular continuation regions. *Biometrics, 39,* 227–236.

## "Sequential Tests of Statistical Hypotheses"

Most statistical work builds on a well-established paradigm. Abraham Wald fully developed and established a new paradigm based on sequential tests of statistical hypotheses. Although formulated in 1943, the new paradigm was first published as "Sequential Tests of Statistical Hypotheses" in *The Annals of Mathematical Statistics* in 1945. In traditional statistical hypothesis testing, an experiment is performed using a predetermined sample size chosen to ensure sufficient statistical power so the null hypothesis is likely to be rejected when in truth it should be rejected. In sequential hypothesis testing, experiments are designed so that after each observation, a decision could be made to accept or reject the null hypothesis or to gather another

observation. Thus, a final decision would be based on a sequence of statistical tests.

Wald's early work in mathematics focused on geometry, but he became a leading econometrician in Vienna. After the German occupation of Austria, Wald escaped to the United States in 1938. He was appointed a fellow by the Carnegie Corporation, studied statistics at Columbia University with Harold Hotelling, and then joined the faculty.

In 1943, while working as part of the Statistical Research Group at Columbia University, the U.S. Navy Department of Ordinance asked for help with a statistical quality control problem. Parts, including munitions, needed to be machined to a certain tolerance or the resulting equipment might not work or might wear out prematurely. It was too expensive and time-consuming to measure precisely every component, so a random sample of components was usually selected and measured to determine whether the manufacturing process was likely to lead to acceptable quality. If tolerances were not met, the manufacturing equipment or processes needed to be modified. The Columbia Statistical Research Group was asked to think about improving the efficiency of the statistical quality control analyses.

There are simple, straightforward ways to improve the efficiency of sampling. For example, you are testing 1,000 units and you are willing to reject the null hypothesis if 50 units are outside of the acceptable tolerance band: If you test 900 units and 50 units are already outside of tolerance, there is no need to test the remaining 100 units. You know you will reject the null hypothesis. Previous researchers, who were cited in Wald's article, identified special circumstances where somewhat more sophisticated versions of sequential hypothesis testing showed greater efficiency than classic hypothesis testing approaches.

Wald realized that costs are associated with rejecting the null hypothesis when it is true, not rejecting the null hypothesis when it is false, and collecting and analyzing incremental data. These costs could be translated into weights and applied to the sequential collection of data and testing of hypotheses. Applying this approach to quality control of war material saved money but perhaps more importantly reduced the time required to get equipment to troops.

Wald realized there were many different sequential test procedures that could be implemented—but how to choose among them was the question. Wald developed a detailed framework and the sequential probability ratio test, which he proved had optimal statistical characteristics including great efficiency. He showed that compared with the most statistically powerful test from a fixed sample size experimental approach, the sequential probability test achieved comparable results with about half the expected sample size. Moreover, he showed the approach could be carried out without knowledge of the probability distribution of the test statistic.

The simplicity and power of Wald's approach was recognized as so important to the war effort that Wald's original papers were restricted so the Axis nations would not be able to use this approach. Thus, the work he did in 1943 was not published until after the end of World War II.

Wald's 1945 paper was divided into an introduction and history and two substantive parts. The first part explored the use of a sequential test comparing a simple hypothesis against a single alternative. The second part expanded his approach to cover a simple or complex hypothesis against a set of alternatives.

Wald died prematurely at age 48 in an airplane crash while he was in route to a lecture tour as a guest of the Indian government. His work, including this paper, went on to influence many statisticians and psychometricians and is the basis of one form of computerized adaptive testing.

*Neal Kingston*

***See also*** Computerized Adaptive Testing

**Further Readings**

Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics, 16,* 117–186.

## Serial Correlation

Serial correlation, or autocorrelation, is defined as the correlation of a variable with itself over successive observations. It often exists when the order of

observations matters, the typical scenario of which is when the same variable is measured on the same participant repeatedly over time. For example, serial correlation is an important issue to consider in any longitudinal designs.

Serial correlation has mainly been considered in multiple regression and time-series models. Multiple regression models are designed for independent observations, where the existence of serial correlation is undesirable. So the main focus in multiple regression is on testing whether serial correlation exists. Conversely, the purpose of time-series analysis is to model the serial correlation to understand the nature of time dependence in the data. The pattern of serial correlation is essential for identifying the appropriate model. This presentation on serial correlation is around regression and time series.

## Multiple Regression Model

Let the multiple regression model be

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i, \quad i = 1, 2, \ldots n, \tag{1}$$

where $y_i$ is the response and $\mathbf{x}_i$ is a $1 \times (k + 1)$ vector consisting of a 1 and the values of the $k$ predictors in the $i$th observation. The assumptions of this model are (a) the expectation of the error $\varepsilon_i$ is 0; (b) $\varepsilon_i$ has constant variance $\sigma^2$; and (c) $\varepsilon_i$ and $\varepsilon_j$ are uncorrelated if $i \neq j$. Let $\hat{\boldsymbol{\beta}}$ be the least squares (LS) estimator of $\boldsymbol{\beta}$ and *SSE* be the sum of squared errors. When the assumptions are valid, $\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimator and $s^2 = SSE/(n - k - 1)$ is an unbiased estimator of $\sigma^2$. When the errors are correlated (violating assumption c), although $\hat{\boldsymbol{\beta}}$ is still unbiased, $s^2$ and the estimated standard error of $\hat{\boldsymbol{\beta}}$ are biased. Consequently, the $F$ or $t$ statistic in testing the significance of $\hat{\boldsymbol{\beta}}$ is misleading. Therefore, it is important to test for the presence of serial correlation. The most widely used test is the *Durbin–Watson d test*, which tests for first-order serial correlation $\rho = \text{Corr}(\varepsilon_i, \varepsilon_{i-1})$ using the autoregressive model

$$\varepsilon_i = \eta\varepsilon_{i-1} + e_i, \quad -1 < \eta < 1. \tag{2}$$

Obviously, $\rho = 0$ if $\eta = 0$. To test the null hypothesis $H_0 : \eta = 0$, the test statistic $d$ is formulated as

$$d = \sum_{i=2}^{n} (\hat{\varepsilon}_i - \hat{\varepsilon}_{i-1})^2 / \sum_{i=1}^{n} \hat{\varepsilon}_i^2,$$

where $\hat{\varepsilon}_i$ are the LS residuals with fitting Equation 1. Let $\hat{\eta} = \sum_{i=2}^{n} \hat{\varepsilon}_i\hat{\varepsilon}_{i-1} / \sum_{i=2}^{n} \hat{\varepsilon}_{i-1}^2$ be the LS estimate of $\eta$ in Equation 2. It follows from $(\hat{\varepsilon}_i - \hat{\varepsilon}_{i-1})^2 = \hat{\varepsilon}_i^2 - 2\hat{\varepsilon}_i\hat{\varepsilon}_{i-1} + \hat{\varepsilon}_{i-1}^2$ that $d \approx 2(1-\hat{\eta})$. Thus, if there is no serial correlation, $d \approx 2$; if the serial correlation is close to 1, $d \approx 0$; and if the serial correlation is close to $-1$, $d \approx 4$. The critical values of $d$ (denote the lower bound as $d_L$ and the upper bound as $d_U$) depend on $n$, $k$, and the significance level of the test. Tables of $d_U$ and $d_L$ can be found in the Appendix of Arnold Studenmund's book. The appropriate decision rules of testing for positive serial correlation are

| | |
|---|---|
| reject $H_0$ | if $d < d_L$; |
| do not reject $H_0$ | if $d > d_U$; |
| inconclusive | otherwise. |

The decision rules of testing for negative serial correlation are

| | |
|---|---|
| reject $H_0$ | if $d > 4 - d_L$; |
| do not reject $H_0$ | if $< 4 - d_U$; |
| inconclusive | otherwise. |

The inconclusive region is one main disadvantage of the Durbin–Watson $d$ test. Moreover, the test ignores serial correlation beyond the first order. It does not allow earlier observed $y$ to predict later $y$ in the regression model either.

The *Lagrange multiplier test* (LM) can overcome the previously mentioned three limitations of the Durbin–Watson $d$ test. To test the null hypothesis that $\rho_j = \text{Corr}(\varepsilon_i, \varepsilon_{i+j}) = 0$ $(1 \leq j \leq q)$ in Equation 1, where $\mathbf{x}_i$ might contain observations $(y_{i-1}, y_{i-2}, \ldots)$, the first step is to obtain $\hat{\varepsilon}_i$ using LS. Then, based on $\hat{\varepsilon}_i$, a new regression equation,

$$\begin{aligned} \hat{\varepsilon}_i &= \mathbf{x}_i\boldsymbol{\theta} + \eta_1\hat{\varepsilon}_{i-1} + \cdots + \eta_q\hat{\varepsilon}_{i-q} + e_i, \\ & i = q + 1, \ldots, n, \end{aligned} \tag{3}$$

is formed. Let $R^2$ be the coefficient of determination (i.e., $R$-square) of fitting Equation 3 using LS; the LM test statistic is

$$\mathrm{LM} = nR^2.$$

The statistic LM is compared with $X_q^2$ to test for significance. If it is significant, the null hypothesis ($\rho_1 = \rho_2 = \cdots = \rho_q = 0$) is rejected. The LS estimates of regression coefficients are biased before the serial correlation among $\varepsilon_i$ is explicitly modeled, which is the focus of the time-series analysis described in the next section.

## Time-Series Analysis

A univariate discrete time series $\{X_t, t = 0, \pm 1, \pm 2, \ldots\}$ is said to be *weakly stationary* if $\mu = E(X_t)$, $\sigma^2 = \mathrm{Var}(X_t)$ and

$$\gamma(h) = Cov(X_t, X_{t+h})$$

do not depend on $t$. As a function of $h$, $\gamma(h)$ is called *autocovariance function* (ACVF). Obviously, $\sigma^2 = \gamma(0)$. For a weakly stationary time series, the serial correlation among $X_t$,

$$\rho(h) = \gamma(h)/\gamma(0) \tag{4}$$

is commonly called the *autocorrelation function* (ACF). It satisfies $\rho(h) = \rho(-h)$, where $h$ is referred to as time lag. Another important concept related to serial correlation is the *partial autocorrelation function* (PACF). It measures the correlation between $X_t$ and $X_{t+h}$ given $X_{t+1} \ldots X_{t+h-1}$. Let $\boldsymbol{\gamma}_h = [\gamma(1), \gamma(2), \ldots, \gamma(h)]'$ be the vector of ACVF and

$$\boldsymbol{\Gamma}_h = \begin{bmatrix} \gamma(0) & \gamma(1) & \ldots & \gamma(h-1) \\ \gamma(1) & \gamma(0) & \ldots & \gamma(h-2) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma(h-1) & \gamma(h-2) & \ldots & \gamma(0) \end{bmatrix}$$

be the matrix of ACVF; the PACF is defined as

$$\alpha(h) = \begin{cases} 1, & h = 0; \\ \phi_{hh}, & h \geq 1, \end{cases} \tag{5}$$

where $\phi_{hh}$ is the last component of the $h \times 1$ vector $\boldsymbol{\phi}_h = \boldsymbol{\Gamma}_h^{-1}\boldsymbol{\gamma}_h$.

In time-series analysis, $\gamma(h)$, $\rho(h)$, and $\alpha(h)$ facilitate the identification of the appropriate model, which is discussed later in this entry.

### *Point Estimators of ACVF, ACF, and PACF*

Based on a sample time series $x_1, \ldots, x_n$, ACVF, ACF, and PACF can be estimated. The recommended estimator of $\gamma(h)$ is given as

$$\hat{\gamma}(h) = n^{-1} \sum_{t=1}^{n-h} (x_{t+h} - \overline{x})(x_t - \overline{x}), \tag{6}$$

where $\overline{x} = \sum_{t=1}^{n} x_t/n$. The denominator in Equation 6 is $n$ instead of $n - h$, which makes $\hat{\boldsymbol{\Gamma}}_h$ nonnegative definite. However, the estimators with either $n$ or $n - h$ as denominator are biased. The bias is small when $n$ is large.

Consistent estimators of $\rho(h)$ and $\alpha(h)$ can be obtained by replacing $\gamma(h)$ with $\hat{\gamma}(h)$ in Equations 4 and 5, respectively.

### *Standard Errors of ACF and PACF Estimates*

To judge the significance of $\hat{\rho}(h)$ or $\hat{\alpha}(h)$, consistent estimates of their standard errors (*SE*s) are needed. Under standard regularity conditions and with a large $n$, $\hat{\boldsymbol{\rho}}_m = [\hat{\rho}(1), \ldots, \hat{\rho}(m)]'$ approximately follows a multivariate normal distribution with mean vector $\boldsymbol{\rho}_m$ and covariance matrix $n^{-1}\mathbf{W}$. The $(i,j)$'s element of $\mathbf{W}$ is given by the Bartlett's formula

$$\begin{aligned} w_{ij} = \sum_{k=-\infty}^{\infty} \{ & \rho(k+i)\rho(k+j) + \rho(k-i)\rho(k+j) \\ & + 2\rho(i)\rho(j)\rho^2(k) - 2\rho(i)\rho(k)\rho(k+j) \\ & - 2\rho(j)\rho(k)\rho(k+i)\}. \end{aligned} \tag{7}$$

The *SE* of $\hat{\rho}(h)$ can be calculated from the diagonal elements of $\mathbf{W}$. When $i = j = h$, Equation 7 reduces to

$$\begin{aligned} w_{hh} = \sum_{k=-\infty}^{\infty} \{ & \rho^2(k+h) + \rho(k-h)\rho(k+h) \\ & + 2\rho^2(h)\rho^2(k) - 4\rho(h)\rho(k)\rho(k+h)\}. \end{aligned} \tag{8}$$

The coefficients $\rho(h)$ and $\alpha(h)$ are used to determine the model for the series $\{X_t, t = 0, \pm 1, \pm 2, \ldots\}$. The most widely used model in time-series analysis is the *causal autoregressive and moving average* (ARMA) *model*

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q},$$

denoted as ARMA($p$, $q$), where $Z_t$ are independent and each follows $N(0, \sigma^2)$. The series $Z_t$ is called a *white noise* process. Special cases of ARMA($p$, $q$) are the autoregressive model AR($p$) when all the $\theta$s are 0, and the moving average model MA($q$) when all the $\phi$s are 0. Because of the importance of the ARMA model, the details for obtaining the estimates of $w_{hh}$ in Equation 8 are given below.

For an MA($q$) model, there are only a few nonzero terms in Equation 8. For AR($p$) and ARMA($p$, $q$) models, although none of $\rho(h)$ is exactly 0, when $h$ is large, $\rho(h)$ is close to 0. Using Equation 8 to estimate $w_{hh}$, the range of $k$ is usually set at a relatively large number (e.g., $n/2$) and $\rho(h)$ is replaced with $\hat{\rho}(h)$. After the estimate $\hat{w}_{hh}$ is obtained, the estimated $SE$ of $\hat{\rho}(h)$ is

$$\widehat{SE}_{\hat{\rho}(h)} = \sqrt{\hat{w}_{hh}/n}. \tag{9}$$

For large $n$, $\hat{\alpha}(h)$ ($h \geq p + 1$) is approximately distributed as $N$ (0,1/$n$). Then the $SE$ of $\hat{\alpha}(h)$ is given as

$$\widehat{SE}_{\hat{\alpha}(h)} = 1/\sqrt{n}, \qquad h \geq p + 1. \tag{10}$$

In practice, this allows us to identify the order $p$ of the ARMA model using the significance of $\hat{\alpha}(h)$ when compared against $N$ (0,1/$n$).

### *Applications of ACF and PACF*

In time-series analysis, sample ACF and PACF are used to identify appropriate models. First, they are used to test for white noise. If the null hypothesis is rejected, they are next used for specifying the appropriate model.

#### Test for White Noise

Testing for white noise is important because no model is needed if a time series is just white noise. Series being tested for white noise can be the original observations or residuals after adjusting for a systematic effect. If $X_t$ is white noise, by definition, $\rho(h) = 0$, and $\alpha(h) = 0$, for all $h \geq 1$. Then it follows from Equation 8 that $w_{hh} = 1$ and $\hat{\rho}(h)$ is approximately distributed as $N$ (0,1/$n$). Equivalently, approximately 95% of $\hat{\rho}(h)$ should fall between the bounds $\pm 1.96/\sqrt{n}$. A sample ACF plot, which shows $\hat{\rho}(h)$ as vertical bars at each $h$ and the two bounds as horizontal lines, is routinely reported in time-series analysis. Whether the series is white noise is indicated by visual inspection of the number and the magnitude that $\hat{\rho}(h)$ exceeds the bounds.

Test statistics for $\rho(1) = \rho(2) = \cdots = \rho(m) = 0$ have also been developed. The so-called *portmanteau statistics*

$$Q = n \sum_{j=1}^{m} \hat{\rho}^2(j) \tag{11}$$

approximately follows $\chi^2_m$. A modified version of $Q$, proposed by Greta Ljung and George Box,

$$Q_{LB} = n(n+2) \sum_{j=1}^{m} \hat{\rho}^2(j)/(n-j), \tag{12}$$

better approximates $\chi^2_m$. The applications of Q and $Q_{LB}$ are illustrated using real data in the last section.

#### Specification of ARMA Models

The ACF and PACF for AR($p$), MA($q$), and ARMA($p$,$q$) processes have certain characteristics. To illustrate these characteristics, plots of analytically derived ACF and PACF of two AR(1), one MA(1), one AR(2), and one ARMA(1,1) processes are provided in Figure 1.

For process 1, an AR(1) process with $\phi_1 = 0.8$, its ACF decays exponentially to 0 when $h$ increases and its PACF only has one nonzero spike at $h = 1$; that is, $\alpha(1) = 0.8$. For process 2, an AR(1) process with $\phi_1 = -0.8$, its ACF is a damped sine wave and its PACF also has only one nonzero value at lag 1, which equals $-0.8$. For process 3, an MA(1) process with $\theta_1 = -0.8$, its ACF has only one nonzero value at lag 1, while its PACF decays exponentially to 0. Process 4 is an AR(2) process with $\phi_1 = 0.8$ and $\theta_1 = -0.64$. Its ACF is a damped sine wave, while its PACF has two nonzero spikes. For process 5, an ARMA(1,1) process with $\phi_1 = 0.8$ and $\theta_1 = -0.8$, its ACF shows exponential decay, while its PACF is a damped sine wave. ACF and PACF for more general AR($p$), MA($q$), and ARMA($p$,$q$) processes have also been

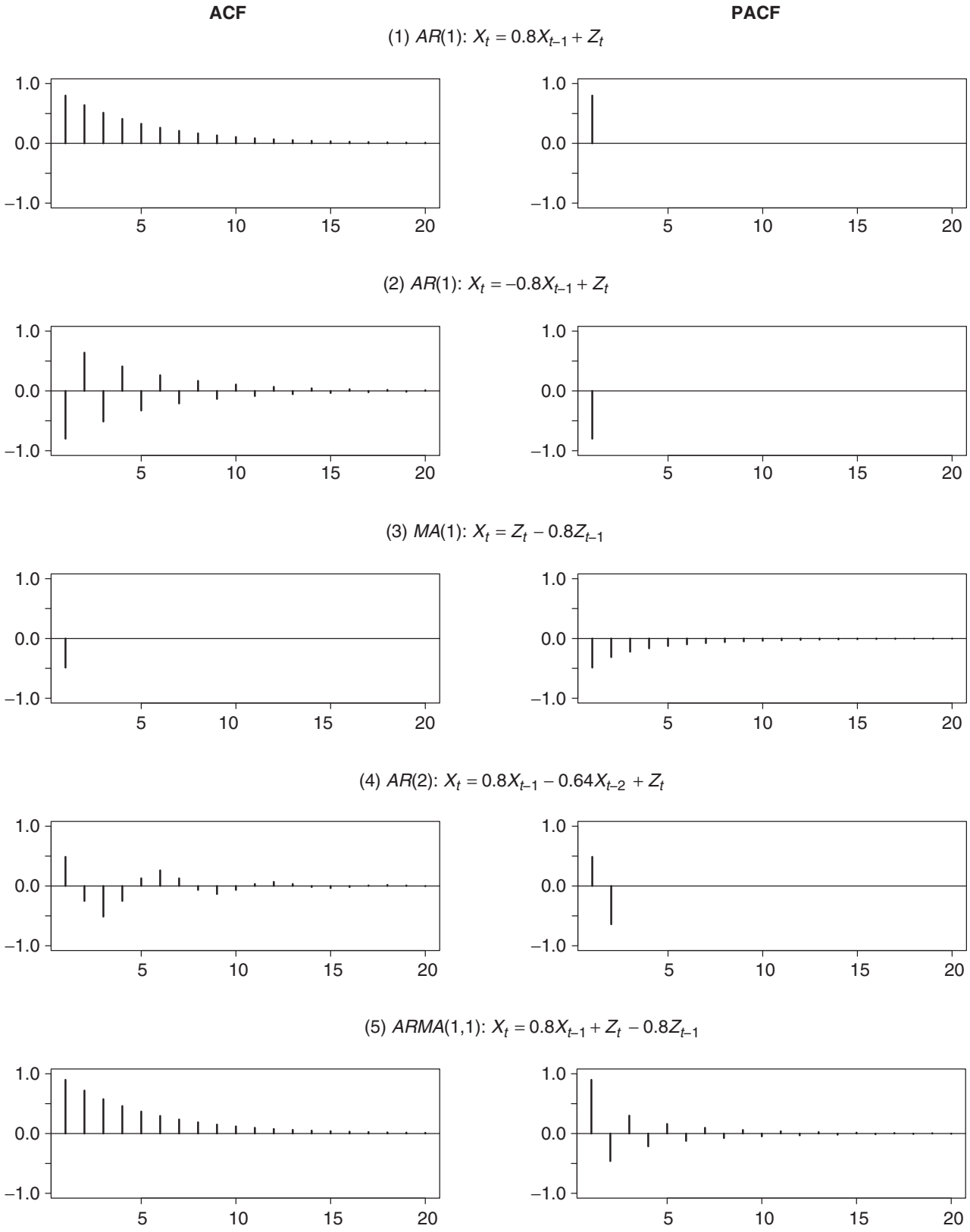



**Figure 1** Plots of ACF and PACF of Five ARMA Processes According to Analytical Formulas

**Table 1** Characteristics of ACF and PACF of AR($p$), MA($q$), and ARMA($p$,$q$) Processes

| *Model* | *ACF* | *PACF* |
|---|---|---|
| AR($p$) | decays exponentially or as a damped sine wave | equals 0 after lag $p$ |
| MA($q$) | equals 0 after lag $q$ | decays exponentially or as a damped sine wave |
| ARMA($p$, $q$) | decays exponentially or as a damped sine wave starting at lag $q$ | decays exponentially or as a damped sine wave starting at lag $p$ |

studied. Their characteristics are summarized in Table 1.

In applications, $\hat{\rho}(h)$ and $\hat{\alpha}(h)$ are calculated and plotted against $h$. Comparison of the sample plots with the characteristics in Table 1 often suggests appropriate models. However, the sample ACF and PACF with real data are rarely as well shaped as those in Figure 1, especially for a series of small number of time points in social sciences. Thus, in addition to ACF and PACF, other criterion such as Akaike information criterion (AIC), or AICC, a bias-corrected version of the AIC, are also used to determine the order of a model.

### *An Illustrative Example*

In this section, the usefulness of ACF and PACF is illustrated by analyzing a real data set. It consists of a 120-day daily score $y_t$ of a schizophrenic patient on a perceptual speed test. This data set was discussed in the book by Gene Glass, Victor Willson, and John Gottman. The raw data were first-order differenced by $x_t = y_t - y_{t-1}$ to achieve stationarity. All the analyses were conducted on $x_t$ using R 2.5.0, a language and environment for statistical computing and graphics.

The sample ACF and PACF plots in Figure 2, with the bounds $\pm 1.96/\sqrt{n}$ for white noise, suggest that the data are not white noise, because $\hat{\rho}(1), \hat{\alpha}(1), \hat{\alpha}(2)$ and $\hat{\alpha}(4)$ exceeded the bounds. The sample PACF decays to 0, whereas the sample ACF has only one big spike, which is at lag 1. These further suggested that an MA(1) model might be appropriate.

Next, a sequence of models from MA(1) to MA(10) was fit to the data, using the function *arima*. In R, *arima* is a function to fit an autoregressive integrated moving average (ARIMA) model to a univariate time series. Both AIC and AICC confirmed MA(1) as the best model. The maximum likelihood estimates of $\theta_1$ and its $SE$ were $\hat{\theta}_1 = -0.688$ and $\widehat{SE} = 0.061$. For an MA(1) process, Equation 8 is simplified to

$$w_{hh} = \begin{cases} 1 - 3\rho^2(1) + 4\rho^4(1), & h = 1; \\ 1 + 2\rho^2(1), & h > 1 . \end{cases} \quad (13)$$

Assuming the true process is $X_t = Z_t - 0.688Z_{t-1}$, $\widehat{SE}_{\hat{\rho}(h)} = \sqrt{\hat{w}_{hh}/n}$ and 95% confidence intervals (CIs) $\hat{\rho}(h) \pm 1.96\sqrt{\hat{w}_{hh}/n}$ are obtained when replacing $\rho(1)$ with $\hat{\rho}(1)$ in Equation 13. In Figure 3, the model implied ACF is plotted as

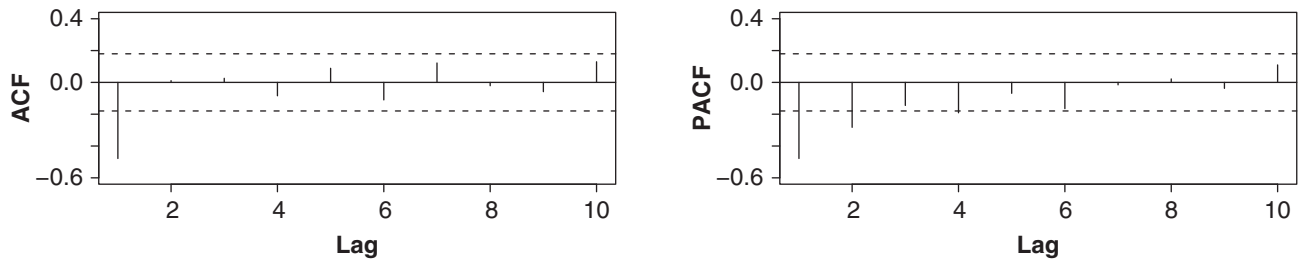



**Figure 2** The Sample ACF (left) and PACF (right) for the Schizophrenia Data and the Bounds $\pm 1.96/\sqrt{n}$ (dashed lines)

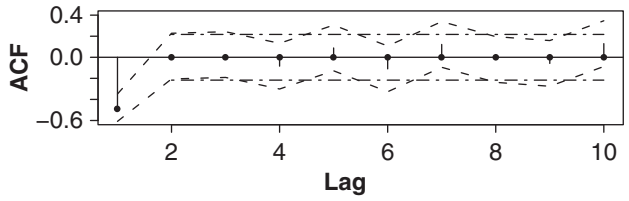



**Figure 3** The Sample ACF (vertical bar), Model Implied ACF (black dot), 95% CI (dashed lines), and the Bounds $1.96\sqrt{\hat{w}_{hh}/n}$ $(h \geq 2)$ (dot-dashed lines) for the Schizophrenia Data

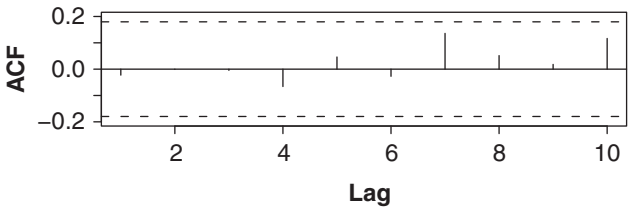



**Figure 4** The Sample ACF (vertical bar) of Residuals from Fitting the MA(1) Model to the Schizophrenia Data and the Bounds $\pm 1.96/\sqrt{n}$ (dashed lines)

black dots, the sample estimates as vertical bars, 95% CI as dashed lines, and the bounds $1.96\sqrt{\hat{w}_{hh}/n}$ $(h \geq 2)$ as dot-dashed lines. The estimated value $\hat{\rho}(1) = -0.478$ was close to the model implied value $-0.467$; the 95% CI always contained the model implied ACF; and no $\hat{\rho}(h)(h \geq 2)$ was outside the bounds. These all suggested the adequacy of the MA(1) model. In practice, because the sample estimates are subject to sampling errors, especially when the time-series length is short, $\hat{\rho}(h)$ is often compared with $\pm 1.96/\sqrt{n}$ for significance, which are bounds for a white noise process.

Finally, for diagnostic purposes, the residuals from fitting the MA(1) model were tested for white noise. In the sample ACF plot of residuals in Figure 4, no $\hat{\rho}(h)$ exceeded the bounds. With $m = 10$, the portmanteau test $Q = 5.086$, $p = 0.885$, and $Q_{\mathrm{LB}} = 5.522$, $p = 0.854$, both failed to reject the null hypothesis that the residual series is white noise. Thus, no further analysis was needed.

*Jiyun Zu and Ke-Hai Yuan*

***See also*** Bivariate Regression; Multiple Regression; Predictor Variable; Time-Lag Study; Time-Series Study; White Noise

**Further Readings**

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (1994). *Time series analysis: Forecasting and control* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Brockwell, P. J., & Davis, R. A. (1991). *Time series: Theory and methods* (2nd ed.). New York: Springer-Verlag.

Brockwell, P. J., & Davis, R. A. (2002). *Introduction to time series and forecasting* (2nd ed.). New York: Springer.

Glass, G. V., Willson, V. L., & Gottman, J. M. (1975). *Design and analysis of time-series experiments*. Boulder: Colorado Associated University Press.

Ljung, G. M., & Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika, 65,* 297–303.

Studenmund, A. H. (2001). *Using econometrics: A practical guide* (4th ed.). Boston: Addison Wesley Longman.

# SHRINKAGE

*Shrinkage* reflects the bias found between sample statistics and inferred population parameters. Multiple regression generally overestimates population values from sample multiple correlation coefficients ($R$) and coefficients of multiple determination ($R^2$). A common adjustment method for overinflation is to use the shrunken or adjusted $R^2$. The adjusted $R^2$ accounts for the amount of shrinkage between the sample $R^2$ and the population squared multiple correlation ($\rho^2$). Similarly, results from a model fitted in one sample are often an overestimate of how it would fit using a separate sample from the same population (i.e., a cross-validation sample), and such results also often need to be adjusted for shrinkage.

This entry begins by explaining why regression overestimates the population parameters. Next, the entry provides an example of shrinkage and discusses its use in cross-validity. This entry ends with a brief discussion of subsequent knowledge in this area.

## Why Regression Overestimates Population Parameters

When working with a random sample of data from a larger population, it is expected that the mean will not exactly match the true population value. Sometimes the mean might be a little higher, and sometimes it might be a little lower. This fluctuation is generally considered to be caused by sampling error. Sampling errors also are present when estimating other numbers, including regression parameters. Regression analyses (in fact all analyses that use the least squares solution) do not account for these positive and negative fluctuations from the true population value when computing the multiple correlation coefficient ($R$). Multiple $R$ is the product-moment correlation between the dependent variable and a linear combination of the set of independent variables. Because least squares maximizes the correlation between the set of independent variables and the dependent variable, and because multiple $R$ cannot be negative (thus all chance fluctuations are in the positive direction), $R$ is overfitted to the sample from which it was estimated. Each sample has its own idiosyncratic characteristics, and ordinary least squares capitalize on these, thus inflating the estimate of $R$.

Increasing the number of predictors also results in an artificially higher multiple $R$ value. If all the chance fluctuations are positive, then adding a variable into the model might increase the multiple $R$ by sampling error variance alone—a situation typically referred to as capitalization on chance. An overestimate of multiple $R$ leads to an overestimate of the coefficient of multiple determination ($R^2$)—an estimate of the proportion of the variance of the dependent variable accounted for by the predictor variables. This positive inflation can easily be found in any statistical program: Add more predictor variables—even ones not significantly related to the dependent variable—and watch both $R$ and $R^2$ increase. Positive inflation becomes even greater for small samples. Because of this bias, statisticians recommend estimating the amount of shrinkage that would occur and adjust the $R^2$ appropriately. In many computer programs, this more appropriate measure of the population value is labeled "adjusted $R^2$."

## Example

Following is a common $R^2$ shrinkage adjustment formula that is often used in statistical packages:

$$R^2_{\text{adjusted}} = 1 - (1 - R^2)\frac{n-1}{n-k-1},$$

where $n$ = sample size and $k$ = number of independent variables.

Consider a model with an $R^2$ of .30 that has three independent variables and a sample size of 100. Using the formula

$$\begin{aligned} R^2_{\text{adjusted}} &= 1 - (1 - .30)\frac{100-1}{100-3-1} \\ &= 1 - .721875 = .278125. \end{aligned}$$

The difference between the $R^2$ and the adjusted $R^2$ is the amount of shrinkage ($.30 - .278125 = .021875$).

The adjusted $R^2$ can vary depending on the size of the $R^2$, the sample size, and the number of independent variables; generally, the smaller the $R^2$ the larger the shrinkage. If the $R^2 = .10$, then the adjusted $R^2$ would be .071875 and the shrinkage is .028. Or, conversely, if the $R^2 = .8$, then the adjusted $R^2$ would be .79375 and the shrinkage is .006.

Additionally, the larger the ratio of the number of independent variables to sample size, the larger the shrinkage. If the number of independent variables in the original example increases from three to ten, the adjusted $R^2$ drops from .278 to .221. Instead, if the sample size drops from 100 to 30, then the adjusted $R^2$ drops from .278 to .219. And if both the number of independent variables increased to ten and the sample size decreased to 30, then the adjusted $R^2$ dramatically declines from .278 to $-.068$ (generally reported as an $R^2$ of zero). As can be observed, when sample size is small, it is wise to choose carefully which variables to include in the model.

## Cross-Validation

The previous section refers to the amount of shrinkage that occurs between the sample value ($R$ or $R^2$) and the population value ($\rho$ or $\rho^2$). However, shrinkage can also occur between one

sample drawn from a population and a second sample drawn from the same population. In other words, shrinkage can occur between the sample multiple $R$ and the population cross-validity multiple correlation ($\rho_c$), or the sample $R^2$ and the squared population cross-validity coefficient ($\rho_c^2$). Simulation studies have shown, however, that it is not appropriate to use the adjusted $R^2$ formula given previously to estimate the squared population cross-validity.

It is common for applied researchers to develop a model and then use it for future prediction. One major problem in prediction is that the weights used in one sample might not generalize to another sample—they are affected by sampling variability. The regression equation might work well in one sample, explaining a significant amount of variance. When applied to a new sample, however, the $R^2$ will often be smaller. Therefore, when using prediction equations, one should cross-validate the equation. Cross-validation applies the weights to a new sample and estimates the $R^2$. If the variance explained is sufficiently large (as defined by the researcher), then the weights are sufficiently generalizable. In a general population, cross-validity ($\rho_c$ or $\rho_c^2$) has been estimated either through empirical means or through formula-based methods.

Splitting a total sample into two subsamples is one example of an empirical method. A regression equation formed in one sample is then used to predict the criterion in the other sample. The biggest problem with this approach is that by splitting the sample, one loses power and generally the standard errors increase, thereby leading to less stable estimates.

The formula-based approach is often preferred because it uses the full sample that provides the most stable regression weights. This approach uses a formula to estimate what the $R^2$ would be in a different sample. Several different formulas have been proposed (see further readings presented at the end of this article for literature that presents these formulas), and similar to the formula for adjusted $R^2$, one would plug in the sample size, number of predictors, $R^2$, and other information obtained through the original regression into the formula. The most appropriate formula to obtain the population squared cross-validity ($\rho_c^2$) often depends on the sample size, number of predictors, and relationships between the predictors in the hypothesized model.

## Further Knowledge

Every sample has unique characteristics, which are often purely a result of chance fluctuations. Models fit to these samples overestimate the amount of variance explained by the independent variables. As such, downward adjustments are often needed. To estimate the true population correlation more accurately and adjust for this capitalization on chance, it is wise to estimate the amount of shrinkage that would occur between the sample and the population parameter. The formula given previously to estimate the adjusted $R^2$ for the population squared multiple correlation ($\rho^2$) is commonly used, and this adjusted $R^2$ is printed in many statistical programs. There seems to be less agreement, however, on the correct formula for estimating the population squared cross-validity ($\rho_c^2$). As the technology continues to improve, it is increasingly likely that new information will be available to guide one's decision. Techniques such as bootstrapping, jackknifing, and simulation studies will add to the knowledge of how and when these different methods and techniques work best. What is fairly consistent, however, is the fact that the larger the original sample, the more robust the estimates and the less shrinkage one would expect.

*Jeffrey Stuewig*

***See also*** Bivariate Regression; Coefficients of Correlation, Alienation, and Determination; Cross-Validation; Pearson Product-Moment Correlation Coefficient; $R^2$; Regression Coefficient; Sampling Error

### Further Readings

Cattin, P. (1980). Estimation of a regression model. *Journal of Applied Psychology, 65,* 407–414.

Cohen, J. C., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Kromrey, J. D., & Hines, C. V. (1995). Use of empirical estimates of shrinkage in multiple regression: A caution. *Educational and Psychological Measurement, 55,* 901–925.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Lawrence Erlbaum.

Raju, N. S., Bilgic, R., Edwards, J. E., & Fleer, P. F. (1997). Methodology review: Estimate of population validity and cross-validity, and the use of equal weights in prediction. *Applied Psychological Measurement, 21,* 291–306.

Raju, N. S., Bilgic, R., Edwards, J. E., & Fleer, P. F. (1999). Accuracy of population validity and cross-validity estimation: An empirical comparison of formula-based, traditional empirical, and equal weights procedures. *Applied Psychological Measurement, 23,* 99–115.

Shieh, G. (2008). Improved shrinkage estimation of squared multiple correlation coefficient and squared cross-validity coefficient. *Organizational Research Methods, 11,* 387–407.

# Significance, Statistical

Statistical significance refers to the difference between two measurements that results from more than randomness. Every research project, experiment, or study that counts, measures, quantifies, or otherwise collects or handles data will ultimately have the need to make comparisons between their data and some other standard. If a difference between two measurements or a measurement and some standard is detected, then that difference is a statistically significant difference if it is caused by an actual difference between the two, rather than simply a result of random variation.

For example, a psychologist might be interested in determining which of two treatments is more effective in treating depression. Having determined some means of determining each treatments' effectiveness, the researcher might, then, administer the different treatments to selected individuals as a part of a designed experiment. Of course, any measure of the treatments' effectiveness, no matter how objective, will only be an estimate of the actual effect. This estimate will vary from the true amount because of many factors, including idiosyncrasies in the testing process, errors in subjective judgment, flaws in measurement, or any number of other sources. Because of the inherent variability in the estimation of any unknown population parameter (in this case, the true effectiveness of each test), in comparing these two measures, the researcher must determine whether the difference between the two results is only caused by variability in the estimation process or if it is caused by an actual difference between the measurements. If the latter is true, then the difference in the measurements is said to be statistically significant. Although there are other methods, this determination is made most frequently by means of hypothesis testing. After briefly discussing the history, this entry discusses hypothesis testing and multiple comparisons and then the objections to statistical significance testing.

## History

Although the formal development of hypothesis testing would not begin until 1925, less formal, ad hoc testing for statistical significance was being done around the turn of the 20th century. In 1908, William Gosset, who is commonly known as "Student," developed his $t$ test for the mean of a normally distributed population with unknown population standard deviation, and before that, in 1892, Karl Pearson published work on chi-square tests for significance with frequency distributions. Perhaps the earliest example of testing for statistical significance is a paper entitled "An Argument for Divine Providence Taken From the Constant Regularity of the Births of Both Sexes" by John Arbuthnot written in 1710, in which he examines birth records in London and concludes that there is good reason to think that the birth rate of males was higher than that of females (i.e., significantly higher). It was not until 1925, however, that R.A. Fisher began the formal development of testing for statistical significance. His work, along with that of Jerzy Neyman and Egon Pearson a few years later, is the foundation for what is known today as hypothesis testing.

## Hypothesis Testing

Fisher, and others writing on this topic at that time, were influenced by the view, which was largely advanced by Karl Popper, that scientific theories must be falsifiable. To that end, the chief purpose of hypothesis testing is not to determine the actual size of the difference between two measurements, but rather to demonstrate that the

difference exists (i.e., is not zero) given some observed data. Specifically, hypothesis testing requires two hypotheses: the null hypothesis (often written $H_0$) and the alternative hypothesis (often written $H_a$ or $H_1$). The null hypothesis is a straw man. It is the theory that the researcher is attempting to falsify by experimentation. The alternative hypothesis is a statement of what the researcher believes to be the true state of affairs. For instance, if a sociologist performs research to determine whether after-school programs reduce the likelihood that participants will be involved in violent crime, the appropriate null hypothesis is that such programs do not reduce the likelihood that participants will be involved in violent crime, whereas one alternative hypothesis might be that these programs do, in fact, reduce such crimes. An educational researcher might want to determine whether preschool attendance increases test scores in at-risk children. That researcher's null hypothesis would be that preschool does not increase test scores, whereas the alternative hypothesis might suggest that it does. In practice, the null hypothesis generally involves the "equals" sign, whereas the alternative hypothesis employs some sort of inequality.

Typically, two types of alternative hypotheses are used: one-sided and two-sided. Although the null hypothesis states the simple and specific equality that the researcher seeks to disprove, the one-sided alternative hypothesis gives the direction in which the true value differs from the hypothesized value. A one-sided alternative hypothesis can be right-tailed, indicating that the true value of the population parameter under consideration is greater than the value hypothesized in $H_0$, or left-tailed, indicating that the true value is less than the hypothesized value. For example, if a particular null hypothesis states that the true mean of a given population is, say, 5, then the corresponding right-tailed alternative hypothesis would be that the true mean is greater than 5, whereas the corresponding left-tailed hypothesis is that the true mean is less than 5. A two-tailed alternative hypothesis is different only in that it does not indicate direction (e.g., the true mean is not equal to 5). These hypotheses must be chosen before the data are collected. If the researcher allows the data to influence the choice of hypotheses, then the test for statistical significance will lose accuracy.

Hypothesis tests rely on the calculation of a statistic from the observed data and the determination of the distribution of that statistic under the terms of the null hypothesis. Whereas in modern times, a multitude of software packages have been created to perform the calculations required for most hypothesis testing, they have not always been available and they are not necessary to the process. They are merely programmed with the probability distributions of various test statistics under the user-inputted null hypothesis. Regardless of the details of how a test is formed and the means by which it is executed, tests for significance allow a researcher to make a determination regarding statistical significance by means of a $p$ value.

After setting the null and alternative hypotheses, collecting data, choosing the appropriate test, and calculating the appropriate test statistic(s), performing the hypothesis test returns a $p$ value. The $p$ value is a measure of statistical significance. Specifically, the $p$ value is the probability of observing data as extreme or more extreme than what was observed given that the null hypothesis is true. In other words, the $p$ value is a measure of how likely (or unlikely) it is for the collected data to have occurred given that the null hypothesis is true. Consequently, a smaller $p$ value means that if the null hypothesis describes the true state of things that a rare event has occurred. Depending on how small the $p$ value is, the researcher might conclude that it is more likely that the null hypothesis is in error. Conversely, a large $p$ value means that the data collected are more in line with what is stated in the null hypothesis.

Typically, the results of a hypothesis are interpreted in two ways. First, a hypothesis test is an indicator of whether to reject the null hypothesis. Having collected data, the researcher compares that value with a previously calculated critical value. If the observed test statistic exceeds the critical value in the appropriate direction (depending on the direction of the alternative hypothesis), then the decision is to reject the null hypothesis. Otherwise, the researcher fails to reject the null hypothesis. The critical value is the value such that, had the observed test statistic exactly equaled the critical value, its $p$ value would be exactly $\alpha$. Consider Figures 1–3. Let the horizontal axis represent values for the test statistic and the shaded area

under the curve represent the probability of the test statistic exceeding that value. Figures 1–3 represent, respectively, left-tailed, right-tailed, and two-tailed alternative hypotheses. Observed test statistic values that fall within this shaded critical region indicate that the null hypothesis should be rejected. Values that fall outside the shaded region indicate that there is not enough evidence to reject the null hypothesis.

Many contexts do not fit this decision-making understanding of statistical significance. In the life sciences, for instance, researchers often are not trying to make a decision but rather are aiming to advance a particular understanding. In this case, researchers typically understand the $p$ value as the weight of evidence against a particular theory. Although a value for $\alpha$ might not be required for making a decision, one is often chosen (commonly .05) and used as provided previously. According to this understanding of statistical significance, a low $p$ value represents strong evidence against the null hypothesis, whereas a higher $p$ value indicates weaker or no evidence. Oftentimes, research of this sort is ongoing and includes repeated experimentation and repetition to understand a particular state of affairs, disprove a scientific theory, or collect evidence in support of a new theory. In many such cases, researchers rely on so-called Bayesian methods to incorporate prior knowledge resulting from previous experience and experimentation into their understanding of statistical significance. As with traditional significance testing, these Bayesian approaches have their weaknesses.

The reader will note that hypothesis testing is not without the chance for error. In fact, there are two types of errors, which are known as Type I and Type II error, that the researcher must consider when attempting to determine statistical significance. Consider Table 1. A Type I error (i.e., the error made when the researcher rejects the null hypothesis when the null hypothesis is, in fact, true) occurs when the $p$ value resulting from the hypothesis test is less than $\alpha$. Recalling the definition of the $p$ value, a $p$ value of .04 indicates that only 4% of the time will results as extreme or more extreme occur given that the null hypothesis is true. Although 4% of the time might be considered a rare event, it is certainly not an impossible one. For that reason, when the researcher chooses $\alpha$, he or she is choosing the

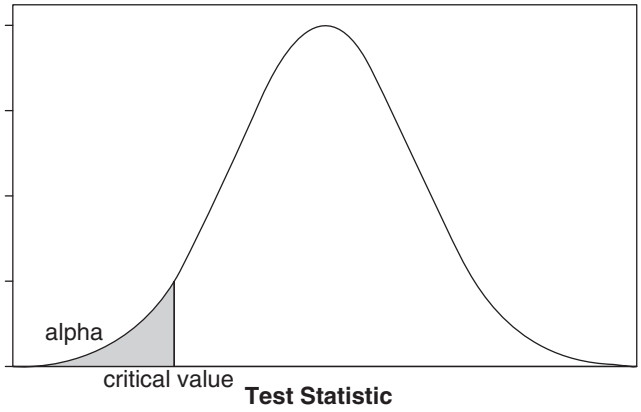



**Figure 1** Critical Region for a Left-Tailed Test

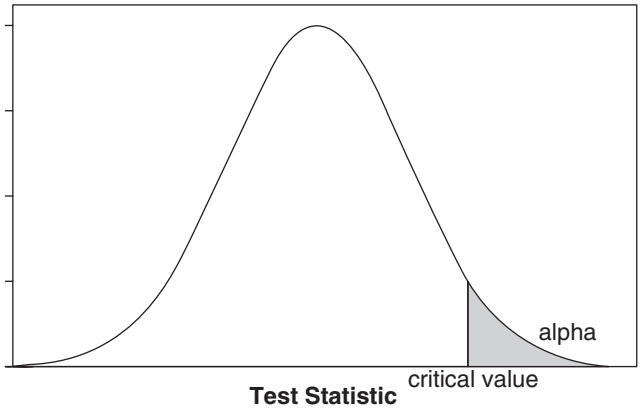



**Figure 2** Critical Region for a Right-Tailed Test

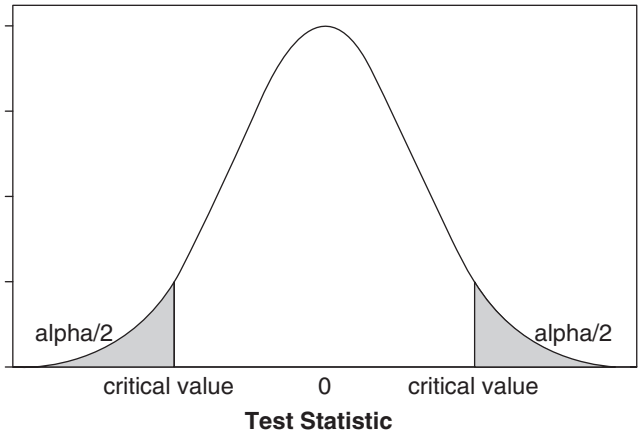



**Figure 3** Critical Region for a Two-Tailed Test

**Table 1** Type I and Type II Errors

| | | *True State of Nature* | |
|---|---|---|---|
| | | *$H_0$ is true* | *$H_0$ is false* |
| *Decision resulting from hypothesis test* | Reject $H_0$ | Type I error | Correct decision |
| | Do not reject $H_0$ | Correct decision | Type II error |

probability that a Type I error will occur. Alpha ($\alpha$) is not set at an exceedingly low value because, all other things held constant, a low probability of making a Type I error results in an increased probability of making a Type II error.

A Type II error occurs when the researcher does not reject the null hypothesis when, in fact, the null hypothesis is false. The probability of making an error of Type II is referred to as $\beta$ and relates to the sensitivity, or power, of a significance test. Power (calculated as $1 - \beta$) is the probability of correctly rejecting the null hypothesis under a specific set of circumstances.

As was mentioned, all other variables held constant, a higher $\alpha$ leads to a lower $\beta$ and vice versa. This makes intuitive sense. Consider a test in which the researcher chooses $\alpha$ very close to zero. In other words, there is almost no chance of rejecting the null hypothesis in error. That is, it will take overwhelming evidence to convince the researcher that the null hypothesis is incorrect. The consequence is that the test is very blunt or insensitive to all but the largest discrepancies between the data and the null hypothesized value. Conversely, suppose a researcher chooses $\beta$ to be exceedingly low (i.e., chooses a test that is very sensitive to even the smallest error). The researcher will be much more likely to call significant differences that might have resulted only from randomness.

Two other variables must be considered when performing a hypothesis test: effect size and sample size. Effect size, which was popularized by Jacob Cohen, refers to the size of difference that the researcher would consider meaningful. For instance, an increase of 3 points on a standardized test might not be of interest to an education researcher, but an increase of 15 points might be practically and theoretically important. This effect size should be fixed at the same time as the hypotheses as well as $\alpha$ and $\beta$.

Finally, the sample size is a function of all these things. Intuitively, one would expect an increased sample size to increase the sensitivity of a test, and rightly so. The researcher should consider the sample size to be a function of the effect size the researcher seeks to observe, $\alpha$, $\beta$, and the hypotheses. Consequently, some studies might be infeasible because a lack of resources or time prevent the gathering of an adequate sample to achieve the desired power, Type I error rate, and effect size. Also of interest is the cost of making either a Type I error versus the cost of making a Type II error. These considerations and others must be factored into the choosing of these experimental settings.

In short, the algorithm for performing a hypothesis test for statistical significance is as follows:

1. Select an appropriate null and alternative hypothesis
2. Set $\alpha$, $\beta$, and effect size
3. Choose appropriate test and test statistic
4. Compute sample size $n$ required for detecting desired effect size, given $\alpha$ and $\beta$
5. Collect data
6. Compute the value of the test statistic from collected data
7. Compare observed test statistic value to critical value
8. Make a decision either to reject the null hypothesis or to fail to reject the null hypothesis
9. Calculate exact $p$ value to determine weight of evidence against the null hypothesis

## Multiple Comparisons

It is frequently the case that multiple determinations of statistical significance must be made. That

is, if a researcher wants to compare three test statistics, instead of two, then one option might be to make multiple pairwise comparisons. An increase in the number of groups being compared, however, quickly inflates the number of pairwise comparisons that must be made. Although 3 groups require only 3 pairwise comparisons, 4 groups require 6 comparisons, 5 groups require 10 comparisons, and so on. In the case of multiple groups, care must be given to the overall, or experiment-wide, error rate. Whereas the error rate for any one of these pairwise tests might be small, the probability that all the tests result in the right decision decreases with the number of tests preformed. To remedy this difficulty, tests have been developed to compare multiple groups in one statistical test (e.g., analysis of variance, for one). In these tests, a low $p$ value indicates that there is a statistically significant difference among at least two groups. This result is less informative than multiple pairwise comparisons would be. It has the advantage, however, of preserving the experiment-wide error rate. In the case of a statistically significant result, typically, the next step is to perform these multiple tests using some sort of correction to the significance level $\alpha$ to correct for experiment-wide error rate inflation.

One such adjustment, the so-called Bonferroni adjustment, is made by dividing the $\alpha$ level of a test by the number of statistical tests being performed to determine the $\alpha$ level to be used for each statistical test. Using that result for each test guarantees that the experiment-wide error rate will be at most $\alpha$.

## Some Objections to Statistical Significance Testing

Although significance testing is common in most areas of quantifiable academic and applied research, it is not without cautions and criticism. Some have even called for abandoning the use of $p$ values altogether. There are two main objections to the traditional, frequentist use of the $p$ value: The first objection has to do with the effect of sample size. The sensitivity of a statistical test increases with the size of the sample being analyzed. As a result, if a very large sample is analyzed, then a hypothesis test will be able to detect even minute differences between the value in the null hypothesis and the observed data. Because of this increased sensitivity, large sample sizes tend to produce statistically significant results, even if the actual difference between the measured data and null hypothesis is very small. Furthermore, a hypothesis test performed on a small sample that demonstrates a large difference between the hypothesized value and the observed data might result in the same $p$ value that would result from a hypothesis test performed on a very large sample that demonstrates only a tiny difference between the two. Consequently, researchers with large amounts of data have to distinguish between statistically significant and practically significant differences. Although there is ample mathematics for determining statistical significance, there is no clear, unambiguous, or agreed on way to determine what is practically significant.

The second main objection to the use of the $p$ value is in its definition: A $p$ value is the probability of observing a test statistic as extreme or more extreme than what was observed given that the null hypothesis is true. Some argue that the phrase "or more extreme" points to values of the test statistic that have not been observed. In other words, if the null hypothesis is rejected, it is because of how unlikely it would be to observe data that were not, in fact, observed.

These objections have led others to consider alternate methods of evaluating data, including Bayesian approaches that incorporate prior information about the parameter being studied into the analysis. In Bayesian methods, the results of the current experiment will be used to update the prior knowledge. That prior knowledge can then be used in future experimentation. This updating procedure is sometimes used in place of or in addition to traditional hypothesis testing. By this means, rather than specifying a specific value for an unknown parameter in the population of interest, a Bayesian researcher treats that parameter as a random quantity and attaches a probability distribution to it. In this way, rather than comparing the data with a specific postulated value, researchers can make probability statements about the value of a population parameter over a range of values. This approach, of course, has its own difficulties, including the nontransferability of subjective probabilities, the computational challenges,

and the use and selection of noninformative and informative prior probability distributions.

*Greg Michaelson and J. Michael Hardin*

***See also*** Analysis of Variance (ANOVA); Bayes's Theorem; Bonferroni Procedure; Coefficient Alpha; Estimation; Inference: Deductive and Inductive; Multiple Comparison Tests; Nonparametric Statistics; Parametric Statistics; Power; Probability Sampling; *p* Value; Sampling Distributions; *t* Test, Independent Samples; *t* Test, One Sample; *t* Test, Paired Samples

**Further Readings**

Barnett, V. (1982). *Comparative statistical inference* (2nd ed.). New York: Wiley.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.) Hillsdale, NJ: Lawrence Erlbaum.

Dolby, G. R. (1982). The role of statistics in the methodology of the life sciences. *Biometrics, 38,* 1069–1083.

Hacking, I. (1965). *The logic of statistical inference.* Cambridge, UK: Cambridge University Press.

Mayo, D. G. (1985). Behavioristic, evidentialist, and learning models of statistical testing. *Philosophy of Science, 52,* 493–516.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioural sciences.* New York: Wiley.

Press, S. J. (2003). *Subjective and objective Bayesian statistics* (2nd ed.). New York: Wiley.

# Significance Level, Concept of

The concept of significance level originates from the discipline of statistical inference, which might be summarized as the application of the scientific method to either observed or collected data. In this setting, it is assumed that the data might be described by some stochastic model, which means that random variation is associated with the variables being measured.

First, an appropriate model for the data must be determined. Then, two hypotheses, known as the null hypothesis (typically denoted $H_0$) and the alternative hypothesis (typically denoted $H_a$), are formulated. These hypotheses are precisely stated in terms of a parameter for the chosen model, such as a mean, proportion, or standard deviation. The null hypothesis specifies a particular value for the parameter of interest, and the alternative hypothesis considers a range of possibilities different from the specified value. Based on the data, the strength of the evidence against the null hypothesis is assessed, considering the probability of observing the data if, in fact, the null hypothesis is correct. Because it was assumed that the data are the product of a random process, it can never be concluded with perfect certainty that the null hypothesis is right or wrong, because even the most unlikely outcomes might occur at one time or another. The probability that the null hypothesis is incorrectly rejected based on the observed data is called the $p$ value of a statistical test, and such a false rejection is called a *Type I error.* Common practice is to decide to reject null hypotheses if $p$ values fall below a certain threshold, such as .05 or .01, so that the probability of a Type I error is small. These thresholds are known as *significance levels,* and it is common to make statements such as "the test was significant at the .05 level" as an alternative to reporting a precise $p$ value.

As an example, suppose a referee for an upcoming professional football game is responsible for the official coin toss. Before the game, the referee decides to determine whether the coin is "fair" by performing the simple experiment of tossing it ten times and counting the number of heads that are observed. If the coin is fair, then the number of heads should follow a binomial distribution with parameters $n = 10$ (number of tosses) and $p = .5$ (probability of a head on each toss). The referee formulates the null and alternative hypothesis, $H_0 : p = .5$ and $H_a : p \neq .5$. This is known as a *two-sided* alternative, which considers the possibility that the true value of $p$ might be either larger or smaller than the value specified in the null hypothesis. After tossing the coin ten times, the referee observes eight heads and two tails. The mathematical probability of observing an outcome at least as extreme as this one, assuming that the coin is actually fair, turns out to be .11. This means that, if the referee rejects the null hypothesis on the basis of this evidence, then he or she would be incorrect in doing so 11% of the time. Because the $p$ value is not smaller than .05, he or she cannot reject the null hypothesis at this significance level and also cannot reject the null hypothesis at

the less conservative .10 significance level. This does not mean, however, that the coin is fair and that the referee can accept the null hypothesis. It simply means that there is insufficient evidence to reject this hypothesis with a high degree of statistical confidence on the basis of the available data.

The previous example considered a discrete probability distribution, one for which there are a countable number of possible outcomes. In ten coin tosses, the total number of heads that can be observed can be no fewer than 0 and no more than 10, and the exact probability of any of these outcomes can be computed using the mathematical functions associated with the binomial distribution.

Often, however, cases are considered for which an infinite number of outcomes are possible, known as continuous distributions. The most commonly applied continuous distribution is the normal, or Gaussian distribution, which has a bell-shaped curve and is an appropriate model for many naturally occurring phenomena. This distribution is defined by its mean $\mu$, which indicates the center of the curve, and its standard deviation $\sigma$, which is a measure of the spread of the curve about its center. A variation of this distribution, which is known as the Student's $t$ distribution, is employed when the normal distribution is considered to be an appropriate model, but both the mean and the standard deviation are estimated from observed data.

To illustrate, suppose there is a study in which a new compound is being tested as a weight-reducing drug. Fifty individuals are enrolled in a study and their weights are taken prior to beginning the drug and then again after 8 weeks of following the specified regimen. The data of interest for this study are the *differences* for each participant from his or her starting weight, so that positive values indicate losses and negative values indicate gains. Because the pharmaceutical company is interested in demonstrating the effectiveness of their drug, the study leader wishes to show that the mean weight loss in the group of participants is significantly greater than 0. She formulates the hypotheses $H_0 : \mu = 0$ and $H_a : \mu > 0$. She specifies a priori that she will reject the null hypothesis of no effect if the test results are significant at the .05 level, which is the level typically specified forconsideration by the scientific community. She observes that, in her participant group, the average weight loss is 1.2 lbs with standard deviation 4.6 lbs. She constructs her test statistic, and using the appropriate methods, she finds that the $p$ value for this paired $t$ test was .036. Because this value is smaller than her previously specified significance level of .05, she can reject the null hypothesis, publish her findings, and seek U.S. Food and Drug Administration approval for her company's drug.

*Michelle Lacey*

***See also*** Alternative Hypothesis; Mean; Normal Distribution; Null Hypothesis; Scientific Method; Standard Deviation; Type I Error

**Further Readings**

Moore, D. S., & McCabe, G. P. (2005). *Introduction to the practice of statistics* (5th ed.). New York: W.H. Freeman & Company.

Ross, S. M. (2005). *Introductory statistics* (2nd ed.). New York: Elsevier.

Salkind, N. J. (2008). *Statistics for people who (think they) hate statistics* (3rd ed.). Thousand Oaks, CA: Sage.

## Significance Level, Interpretation and Construction

Hypothesis testing is not set up so that a researcher can absolutely prove a null hypothesis. Rather, it is set up so that when a researcher does not find evidence against the null hypothesis, he or she fails to reject the null hypothesis. When the researcher does find strong enough evidence against the null hypothesis, he or she rejects the null hypothesis. This, although often confusing, is a subject with vast field of statistical application. In hypothesis testing, the significance level at some preassigned small value $\alpha$ is used to control the probability of Type I error (rejecting the null hypothesis when it is true), which is vital in building up theory and methods. Among many significance-level-related notions for interpretation and elements for construction of hypothesis testing, some important ones are addressed in this entry.

## The Need of Level of Significance in Hypothesis Testing

In hypothesis testing, it is often impossible, too costly, or too time-consuming to obtain the entire population data on any variable to determine whether a null hypothesis is true. Decisions of hypothesis testing, thus, should be made using sample data. Whenever an experiment in collecting evidence is undertaken, despite how seriously care and controls are introduced by the researcher, the outcome is always subject to some variability of chance. Hence, falsely rejecting the null hypothesis or falsely not rejecting the null hypothesis always exists. The classic way to solve this dilemma is to confine the class of tests for consideration. A conventional setting for confining the tests is assigning the level of significance. Once a test is chosen to deal with a given null hypothesis, the calculated value of the test statistic is compared with tables of critical values at specific level of significance. If the calculated value exceeds the critical value, then the null hypothesis is rejected.

## Setting the Level of Significance

A significance level $\alpha = .05$ means that there is a 5% chance a researcher will accept the alternative hypothesis (reject the null hypothesis) when the null hypothesis is true. Then, in the long run, the proportion of times the researcher will make a Type I error will be .05. However, the selection of the level of significance should not be affected by the results of the data; the researcher should choose it before the data have been collected. If the level $\alpha$ is affected by the data, a bias on stated error probabilities should be entered to the study. In fact, for avoiding a bias of the study, a complete decision process involving the selection of test statistic, the choice of significance level $\alpha$, for determining one-sided or two-sided test, and the cutoff points must be set up in advance of making any observation of the data.

## Size, $p$ Value, and Significance Level

People often do not make a clear distinction between the size and significance level of a test, and sometimes these two terms are used interchangeably. By letting null hypothesis $H_0 : \theta \in \Theta_0$, a clear distinction might be made by saying that a test with power function $\pi(\theta)$ is a size $\alpha$ test if $\sup_{\theta \in \Theta_0} \pi(\theta) = \alpha$ and is a level $\alpha$ test if $\sup_{\theta \in \Theta_0} \pi(\theta) \leq \alpha$. Constructing a size $\alpha$ test sometimes is so difficult for computation that the researcher might turn to setting the compromise of constructing a level $\alpha$ test. Once the level of significance has been set as $\alpha$, the rule of hypothesis testing is to reject the null hypothesis when its size is smaller than $\alpha$.

One useful procedure in hypothesis testing is to look at the total probability of the set of observations, under the model of null hypothesis, which are at least as extreme as observed. This cumulated probability is called the $p$ value of the test. It is also very important to distinguish between $p$ value and significance level. The level of significance is a notion in the context of the approach developed by Jerzy Neyman and Egon Pearson, dealing with the null and alternative hypotheses; and the $p$ value is a notion, introduced by J. D. Gibbons and J. W. Pratt, in the context of significance test, dealing with a null hypothesis without specifying an alternative one. In connection of these two terms, the $p$ value represents the minimum significance level for which the null hypothesis would have been rejected. In this circumstance, if the p value is lower than the significance level $\alpha$, then the null hypothesis is rejected.

## Optimal Significance Level $\alpha$ Test

Suppose that we have a random sample $\mathbf{X} = (X_1, \ldots, X_n)$ drawn from a distribution with pdf $f(x, \theta)$. One of the needs in setting significance level is restricting a class of tests for developing optimal test. Given a significance level $\alpha$ and considering the simple null hypothesis $H_0 : \theta = \theta_0$, the general interest in hypothesis testing is searching a best one from the following class of level $\alpha$ tests,

$$K_\alpha = \left\{ C : P_{\theta_0}\{\mathbf{X} \in C\} \leq \alpha \right\},$$

where each $C$ in $K_\alpha$ represents a level $\alpha$ critical region. For this hypothesis testing problem, Neyman and Pearson proposed to specify a simple alternative hypothesis, saying $H_1 : \theta = \theta_1$ and defined the best one with the largest power, that is, the largest probability of rejecting $H_0$ when $H_1$ is true. Denote the likelihood function as

$L(\theta, \mathbf{x}) = \mathbf{\Pi}_{i=1}^{n} f(x_i, \theta)$ with $\mathbf{x}$ being the observation of random sample $\mathbf{X}$, the most powerful test has a critical region determined by

$$\frac{L(\theta_0, \mathbf{x})}{L(\theta_1, \mathbf{x})} \leq k,$$

where $k$ satisfies $\alpha = P\left(\frac{L(\theta_0, \mathbf{X})}{L(\theta_1, \mathbf{X})} \leq k\right)$. To maximize the power of the test, the sample points that are more likely under $H_1$ than under $H_0$ should be put into the critical region. Hence, if a sample point $\mathbf{x}$ makes the denominator of the likelihood ratio $\frac{L(\theta_0, \mathbf{x})}{L(\theta_1, \mathbf{x})}$ larger, this $\mathbf{x}$ is the point more likely under $H_1$. This is why the requirement uses likelihood ratio to set up level $\alpha$ critical region.

A test considering only one hypothesis such as $H_0 : \theta = \theta_0$ is called a significance test or pure significance test. In fact, significance tests were given their modern justification and then popularized by Ronald Fisher, who derived most of the test statistics that researchers now use, in a series of papers and books during 1920s and 1930s. Unlike the formulation of Neyman and Pearson, there is no optimality theory to support the classic techniques of significance test. It is, therefore, argued that a significance test cannot use likelihood ratio for test derivation. Recently, H. C. Chen defined the best level $\alpha$ significance test as the one in $K_\alpha$ with smallest volume of its acceptance region. To accomplish this, Chen constructed critical region only through an observation's probability $L(\mathbf{x}, \theta_0) = \mathbf{\Pi}_{i=1}^{n} f(x_i, \theta_0)$, the joint pdf of the random sample when $H_0$ is true. By letting $\mathbf{x}_a$ and $\mathbf{x}_b$ be two observation points, observation $\mathbf{x}_a$ is more probable than $\mathbf{x}_b$ if the following holds:

$$\frac{L(\mathbf{x}_a, \theta_0)}{L(\mathbf{x}_b, \theta_0)} \geq 1.$$

To minimize the volume of the test, the sample points of the highest densities have to be set as the acceptance region. Chen shows that the level $\alpha$ significance test with acceptance region of highest densities as

$$A_{hdt} = \{\mathbf{x} : L(\mathbf{x}, \theta_0) \geq a\}, \text{ subjected to}$$
$$1 - \alpha = P_{\theta_0}\{\mathbf{X} \in A_{hdt}\}$$

is the best one.

Let us see an example for interpretation. Consider a random variable X that has a pdf $f(x, \theta)$ under null hypothesis $H_0 : \theta = \theta_0$ as follows:

| $x$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $f(x, \theta_0)$ | 0.7 | 0.025 | 0.025 | 0.20 | 0.05 |

Suppose that the significance level is $\alpha = .05$. The highest density significance test has acceptance region $A_{hdt} = \{x : f(x, \theta_0) \geq a\}$ subjected to $0.95 = P_{\theta_0}\{X \in A_{hdt}\}$, which leads to $A_{hdt} = \{1, 4, 5\}$. The classic significance test chooses test statistic X and sets rejection region $\{x : x \geq b\}$ subjected to $.05 = P_{\theta_0}\{X \geq b\}$. The resulting rejection region is $C_{cla} = \{5\}$ and acceptance region is $A_{cla} = \{1, 2, 3, 4\}$. Let us also consider an alternative hypothesis $H_1 : \theta = \theta_1$, with its corresponding pdf and likelihood ratio being shown as follows:

| $x$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $f(x, \theta_1)$ | 0.07 | 0.025 | 0.025 | 0.08 | 0.80 |
| $f(x, \theta_1)/f(x, \theta_0)$ | 0.1 | 1 | 1 | 0.4 | 16 |

The most powerful test of Neyman–Pearson chooses rejection region based on

$$0.05 = P_{\theta_0}\left\{x : \frac{f(x, \theta_1)}{f(x, \theta_0)} \geq c\right\}$$

that leads to the rejection region $C_{np} = C_{cla}$ and the acceptance region $A_{np} = A_{cla}$.

## Null Hypothesis Setting

Often, major decisions for research of many sciences are based on the results of a hypothesis test. However, in most circumstances, researchers' attitudes to the null hypothesis and the alternative hypothesis are different, and they know that when the probability of Type I error is minimized, the probability of the Type II error must be increased. With this concern, the significance level specifies the criterion for rejecting the null hypothesis. Thus, the decision ultimately rides on the question, "Which hypothesis erroneously rejected is more serious in consequence?" No rule handles the setting of null hypothesis for all the statistical hypothesis testing problems.

There are generally two philosophies in setting the null hypothesis $H_0$. The most commonly accepted rule for setting null hypothesis is through the following philosophy:

> If one wishes to prove that A is true, one first assumes that it is not true.

Suppose that we are studying a new drug treatment for reducing the patient's pain. We might state the null hypothesis $H_0$ as "There will be no significant difference in reducing pain" and the alternative hypothesis $H_1$ as "There will be a significant difference in reducing pain." This philosophy defined the hypotheses based on the concept of proof by contradiction that gives the null hypothesis a statement of "no effect" or "no difference." It, in fact, tries to prevent the consumer's risk with that if we are going to make an error, we would rather error on the side of consumer's benefit. As the example given previously, when it is claiming that a new drug is better than the current drug for treatment of the same symptoms, the null hypothesis indicates that no difference exist between these two drugs.

There is an approach of philosophy that considers a situation employed in some engineering problems. The quality of product is distributed from background noise and abnormal noise. If the process operator adjusts the process based on the tests performed periodically for hypothesis defined from the commonly accepted rule, it will often overreact to the background noise and deteriorate the performance of the process. Hence, to prevent this unnecessary adjustment, a rule guiding the null hypothesis is as follows:

> Unless it is broken, we will not fix it.

This philosophy tries to prevent the manufacturer's risk with that if we are going to make an error, we would rather error on the side of manufacturer's benefit.

One useful technique following this philosophy is the control chart. In the online quality control, without evidence of existing assignable cause, the manufacturing process is considered to be in control. For example, one constructs the $\overline{X}$ chart as

$$\text{UCL} = \mu_0 + 3\frac{\sigma_0}{\sqrt{n}},$$
$$\text{LCL} = \mu_0 - 3\frac{\sigma_0}{\sqrt{n}},$$

where $\mu_0$ and $\sigma_0$, respectively, are the mean and standard deviation for in-control process, and the process is considered to be in control when sample mean $\overline{x}$ falls within the control limits, lower control limit (LCL) and upper control limit (UCL). In a situation where $\mu_0$ and $\sigma_0$ are unknown, they might be replaced by estimates computed from the training samples. This chart aims in detecting whether there is a mean shift. So, the hypothesis represented behind this chart is $H_0 : \mu = \mu_0$, whereas a probability 0.9973 of the acceptable region does favor no mean shift. Following this philosophy for preventing the manufacturer's risk, the wish of proving that the process is in control is to set it up as the null hypothesis.

## Two Cautions in Applying Significance Tests

Test results sometimes lead researchers to inappropriate conclusions. Can a researcher say that the data show practical significance, when it reaches statistical significance with significance level $\alpha$ as small as .01? The sample size of the data set directly impacts its corresponding $p$ value. On the one hand, large sample sizes do produce small $p$ values even when there is no effect or differences between groups. In this case, statistical significance is practically not meaningful. On the other hand, a sample size that is too small can result in a failure to identify an effect or a difference when one truly exists. Researchers should always verify the practical relevance of the computed results. Let us use one example for interpretation. Suppose that we have a random sample $X_1, \ldots, X_n$ from a normal distribution $N(\mu, 1)$ and we consider hypothesis $H_0 : \mu = 0$ versus $H_1 : \mu > 0$. We also assume that we have observation $\overline{x} = 0.01$and consider the test rejecting $H_0$ if $\sqrt{n}\overline{X} > z_{0.01} = 2.33$ occurs. If $n = 30$, this observation provides evidence statistically insignificant. However, when $n > 233$, it provides evidence statistically significant but $\mu = 0.01$ is not practically significant, with a difference being from $\mu = 0$ in many applications.

One other caution in the significance test is the tendency for multiple comparisons to yield

artificial significant differences even when the null hypothesis of no differences is true. If we apply test of significance level .05 for a study of 20 multiple comparisons, one comparison will be likely to yield a significant result despite the null hypothesis being true. This problem can be more serious when researchers deal with some modern multiple comparison problems such as the gene expression analysis where more than 10,000 genes need to be simultaneously studied to classify influential genes. As a result of classic *t* test for differences of means, hundreds of noninfluential genes could be identified as influential ones just by chance.

*Lin-An Chen*

***See also*** Clinical Significance; Null Hypothesis; Significance, Statistical; Significance Level, Concept of; Type I Error; Type II Error

**Further Readings**

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh, Scotland: Oliver and Boyd.

Gibbons, J. D., & Pratt, J. W. (1975). p-value: Interpretation and methodology. *The American Statistician, 29,* 20–25.

Lehmann, E. L. (1993). The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? *Journal of the American Statistical Association, 88,* 1242–1249.

Neyman, J., & Pearson, E. S. (1967). *Joint statistical papers*. Cambridge, UK: Cambridge University Press.

Schervish, M. J. (1996). p-values: What they are and what they are not. *The American Statistician, 50,* 203–206.

# Sign Test

The sign test, first introduced by John Arbuthnot in 1710, is a nonparametric test that can be applied in a variety of situations such as the following:

When the data have two possible outcomes, it can be used to test whether these two outcomes have equal probabilities. In this situation, the sign test can be regarded as a special case of the binomial test.

It can be used when the variable of interest is ordinal, interval, or ratio, and one wants to test whether the median of this variable is some given value.

It can be used in paired data analysis where there are two categorical variables and one continuous or ordinal variable. One of the categorical variables has only two values, such as "before treatment" and "after treatment" or "left" and "right," and the other categorical variable identifies the pairs of observations. The directions of change in the continuous or ordinal variable can be increase (denoted as "+"), decrease (denoted as "−"), or no change (denoted as "0"). The sign test is used to test whether the numbers of change in each direction are equal or not; that is, it tests whether there are equal probabilities for values "+" and "−."

McNemar's test, which is a nonparametric test used on categorical data to determine whether the row and column marginal probabilities are equal, can be viewed as a variation of the sign test.

The Cox–Stuart test, which is used to test for the presence of a trend in one sample ordinal, interval or ratio data, can also be regarded as a variation of the sign test.

These situations are discussed in subsequent sections in the same order as they are listed here. In the last section, the sign test is compared with other similar tests.

## Sign Test for One Sample Categorical Data

Suppose that $X_1, X_2, \ldots, X_N$ are the observations on $N$ subjects, and assume the following:

1. The random variables $X_1, X_2, \ldots, X_N$ are mutually independent.
2. Each $X_i$ takes one of three possible values, denoted as "+" (plus), "−" (minus), or "0" (zero).
3. The $X_i$'s have consistent probability distributions, in that if $P_+ > P_-$ for one $i$, then it also holds for the rest indices, where $P_+$ denotes the probability that $S_i =$ "+," $P_-$ denotes the probability that $S_i =$ "−." The same is true for $P_+ < P_-$ and $P_+ = P_-$.

To test whether $X_i$ has equal probabilities of taking "+" and "−," the test statistics to use is

$S_+ =$ the total number of "+"s. The null distribution of $S_+$ is binomial $(n; 1/2)$, where $n$ is the number of nonzero $X_i$s. The following tests can be performed.

*The Two-Sided Test*

$$H_0 : P_+ = P_- \text{ versus } H_1 : P_+ \neq P_-.$$

Denote $s$ as the observed $S_+$, then the $p$ value is $2\min(P(S_+ \leq s), P(S_+ \geq s))$ where

$$P(S_+ \leq s) = \sum_{i=0}^{s} \binom{n}{i} \left(\frac{1}{2}\right)^n$$

and

$$P(S_+ \geq s) = \sum_{i=s}^{n} \binom{n}{i} \left(\frac{1}{2}\right)^n.$$

If $n > 20$, then $Z = \frac{2S_+ - n}{\sqrt{n}}$ is approximately standard normal, and the normal approximation to binomial can be used to calculate the $p$ value as

$$p \text{ value} = 2\left[\min\left(P\left(Y \leq \frac{2s - n - \text{sign}(2s-n)}{\sqrt{n}}\right), P\left(Y \geq \frac{2s - n - \text{sign}(2s-n)}{\sqrt{n}}\right)\right)\right],$$

where $Y$ is a standard normal random variable, and

$$\text{sign}(2s - n) = \begin{cases} 1 & \text{if } 2s > n \\ 0 & \text{if } 2s = n \\ -1 & \text{if } 2s < n \end{cases}$$

*The One-Sided Lower-Tail Test*

$$H_0 : P_+ = P_- \text{ versus } H_1 : P_+ < P_-$$

Denote $s$ as the observed $S_+$, then the $p$ value is $P(S_+ \leq s) = \sum_{i=0}^{s} \binom{n}{i} (\frac{1}{2})^n$.

If $n > 20$, using normal approximation

$$p \text{ value} = P\left(Y \leq \frac{2s - n - \text{sign}(2s-n)}{\sqrt{n}}\right),$$

where $Y$ is a standard normal random variable.

*The One-Sided Upper-Tail Test*

$$H_0 : P_+ = P_- \text{ versus } H_1 : P_+ > P_-.$$

Denote $s$ as the observed $S_+$, then the $p$ value is $P(S_+ \geq s) = \sum_{i=s}^{n} \binom{n}{i} (\frac{1}{2})^n$.

If $n > 20$, using normal approximation

$$p \text{ value} = P\left(Y \geq \frac{2s - n - \text{sign}(2s-n)}{\sqrt{n}}\right),$$

where $Y$ is a standard normal random variable.

**Example**

Suppose that the cholesterol level of 30 patients was measured right before and 10 days after they received some treatment. It was observed that 30 days after they received the treatment, 20 of the 30 patients had a lower cholesterol level and 10 had a higher cholesterol level. To determine whether the treatment significantly lowered the cholesterol level, let "+" denote a higher cholesterol level after the treatment, and "−" denote a lower one; thus, there are 10 "+"s and 20 "−"s in the sample. The test is $H_0 : P_+ = P_-$ vs $H_1 : P_+ \neq P_-$ based on $S_+ = 10$. Because the sample size $n = 30 > 20$, a normal approximation is used, and the $p$ value is calculated as

$$\begin{aligned} p &= 2\min\left(P\left(Y < \frac{2 \times 10 - 30 + 1}{\sqrt{30}}\right), P\left(Y < \frac{2 \times 10 - 30 + 1}{\sqrt{30}}\right)\right) \\ &= 2\min(0.051, 0.049) \\ &= 0.10. \end{aligned}$$

This nonsignificant $p$ value indicates that the cholesterol level was not significantly lowered by the treatment.

## Sign Test for the Median

Suppose there is a sample of continuous or ordinal random variables $Z_1, Z_2, \ldots, Z_n$ from some distribution with unknown median $m$, and one wants to test the null hypothesis $H_0 : m = m_0$ against a one-sided or two-sided alternative hypothesis.

One way to test this is to define new random variables $S_1, S_2, \ldots, S_n$, such that $S_i$ is assigned the value "+" if $Z_i$ is greater than $m_0$, "−" if $Z_i$ is less than $m_0$, and "0" if $Z_i$ equals $m_0$. If $H_0$ is true (i.e., the sample comes from a distribution with the median $m_0$), then on average, there are equal numbers of observations to be greater and smaller than $m_0$. Denote the number of plus signs as $n_+$ and the number of minus signs as $n_-$. Thus, $H_0 : m = m_0$ is equivalent to $H_0 : p_+ = p_-$, where $p_+ = P(S_i = \text{"+"})$, and $p_- = P(S_i = \text{"−"})$. The alternative hypothesis $H_1 : m > m_0$ is equivalent to $H_1 : p_+ > p_-$. The two-sided hypothesis $H_1 : m \neq m_0$ is equivalent to $H_1 : p_+ \neq p_-$. The sign test for one sample categorical data, as discussed in the last section, can be used on $S_1, S_2, \ldots, S_n$ to test the previous hypotheses about $p_+$ and $p_-$ and, equivalently, the corresponding hypotheses about $m$.

## Sign Test for Two Paired Samples

Suppose that there are the following paired observations on each of $n$ subjects:

$$(Y_{11}, Y_{21}), \ldots, (Y_{1n}, Y_{2n}).$$

The two numbers within each pair $(Y_{1i}, Y_{2i})$are compared and a binary variable $Z_i$ is defined based on the comparison as: If $Y_{1i} < Y_{2i}$, then $Z_i =$ "+" and $Z_i =$ "−" if $Y_{1i} > Y_{2i}$. If $Y_{1i} = Y_{2i}$, then $Z_i =$ "0." Assume the following:

- The random vectors $(Y_{1i}, Y_{2i})$, $i = 1, 2, \ldots, n$ are mutually independent.
- Each $Y_{ij}$, $i = 1, 2$, $j = 1, 2, \ldots, n$ should be ordinal [i.e., each pair $(Y_{1i}, Y_{2i})$ can be determined to be "+," "−," or "0"].
- Each pair difference $Y_{2i} - Y_{1i}$, $i = 1, 2, \ldots, n$ comes from a continuous population (not necessarily the same) that has a common median.

The null hypothesis to test is median $(Y_{2i} - Y_{1i}) = 0$, $i = 1, 2, \ldots, n$, that is $P(Z_i =$ "+") $= P(Z_i =$ "−"), meaning that there is not a significant difference between the two pairs. This test is equivalent to the simple sign test on the data $Z_1, Z_2, \ldots, Z_n$.

**Table 1** Days of Pain Relief

| *Patient ID* | $d_A$ : *Days of Pain Relief by Drug A* | $d_B$ : *Days of Pain Relief by Drug B* | *Sign* $(d_A - d_B)$ |
|---|---|---|---|
| 1 | 14 | 15 | − |
| 2 | 25 | 24 | + |
| 3 | 33 | 25 | + |
| 4 | 12 | 7 | + |
| 5 | 9 | 12 | − |
| 6 | 14 | 7 | + |
| 7 | 8 | 5 | + |
| 8 | 4 | 2 | + |
| 9 | 13 | 17 | − |
| 10 | 12 | 11 | + |

### *Example*

Table 1 shows the days of pain relief provided by two analgesic drugs in 10 patients. Is there any evidence that drug A provides longer relief than drug B?

Here the null hypothesis is that the median of the difference is zero. The alternative hypothesis is that median of the difference is not 0. The last column gives the sign of $d_A - d_B$. Notice that there are 7 "+" signs out of 10 patients. The two-sided test has a $p$ value

$$2\min\left(\sum_{i=7}^{10}\binom{10}{i}\left(\frac{1}{2}\right)^{10}, \sum_{i=0}^{7}\binom{10}{i}\left(\frac{1}{2}\right)^{10}\right)$$
$$= 2\min(0.95, 0.17) = 0.34,$$

which is greater than .05, indicating that there is no evidence that drug A provides longer relief than drug B. Note that the Wilcoxon signed rank test is also appropriate in this case, and it is more powerful because it uses the information about the magnitude of the differences as well as the signs.

## McNemar's Test

Suppose that each $Y_{ij}$ in the paired data $(Y_{11}, Y_{21}), (Y_{12}, Y_{22}), \ldots, (Y_{1n}, Y_{2n})$ is nominal with two categories denoted by "0" and "1." The pairs $(Y_{1i}, Y_{2i})$ are mutually independent, and the difference $P(Y_{1i} = 1, Y_{2i} = 0) - P(Y_{1i} = 0, Y_{2i} = 1)$ is negative for all $i$, or zero for all $i$, or positive for all $i$.

**Table 2** Data Table for McNemar's Test

| | $Y_2 = 0$ | $Y_2 = 1$ |
|---|---|---|
| $Y_1 = 0$ | $a$ | $b$ |
| $Y_1 = 1$ | $c$ | $d$ |

The question of interest is, "Is there a difference between the probability of (0, 1) and the probability of (1, 0)?" This question is often asked when $Y_{1i}$ in the pair $(Y_{1i}, Y_{2i})$ represents the status of the subject before the experiment and $Y_{2i}$ represents the status of the same subject after the experiment. Such a question also arises when $Y_{1i}$ in the pair $(Y_{1i}, Y_{2i})$ represents the measurement with an instrument and $Y_{2i}$ represents the measurement with another.

The data are usually summarized in a $2 \times 2$ contingency table as shown in Table 2, where *a, b, c,* and *d* are the numbers of (0,0), (0,1), (1,0) and (1,1), respectively.

The hypotheses are usually expressed as

$$H_0 : P(Y_{1i} = 1, Y_{2i} = 0) = P(Y_{1i} = 0, Y_{2i} = 1) \text{ for all } i$$

versus

$$H_1 : P(Y_{1i} = 1, Y_{2i} = 0) \neq P(Y_{1i} = 0, Y_{2i} = 1) \text{ for all } i.$$

The previous hypothesis is equivalent to

$$H_0 : P(Y_{1i} = 0) = P(Y_{2i} = 0) \text{ for all } i$$

versus

$$H_1 : P(Y_{1i} = 0) \neq P(Y_{2i} = 0) \text{ for all } i,$$

and

$$H_0 : P(Y_{1i} = 1) = P(Y_{2i} = 1) \text{ for all } i$$

versus

$$H_1 : P(Y_{1i} = 1) \neq P(Y_{2i} = 1) \text{ for all } i,$$

so this is the test of whether $Y_{1i}$ and $Y_{2i}$ have the same marginal distributions.

The test statistic for McNemar's test is usually written as

**Table 3** Summary of Test Result

| | *Result Was Positive Before Treatment* | *Result Was Negative Before Treatment* |
|---|---|---|
| Result was positive after treatment | 12 | 5 |
| Result was negative after treatment | 20 | 13 |

$$T_1 = \frac{(b-c)^2}{b+c}.$$

Under the null hypothesis, $T_1$ follows $\chi^2$ distribution with one degree of freedom, that is $T_1 \sim {\chi_1}^2$.

### *Example*

In a study, a test is performed in 50 patients before treatment and after treatment. Before treatment, 32 patients had positive results and 18 had negative results (Table 3). After treatment, 20 of the 32 patients with positive pretreatment test results now had negative results, and 5 of the 18 patients with negative pretreatment results now had positive results. Is there a significant change in the test result before and after treatment?

The test statistic is

$$T = \frac{(5-20)^2}{(5+20)} = 9.$$

The $p$ value is $.003 < .05$, indicating that the treatment can significantly improve the test result.

## Cox and Stuart Test for Trend

The Cox and Stuart test was introduced in 1955 and it is used to test for the presence of a trend. A sequence of numbers is said to have a trend if in the sequence the later numbers tend to be larger than the earlier numbers (upward trend) or smaller than the earlier numbers (downward trend).

Suppose that the data consist of continuous or ordinal observations $X_1, X_2, \ldots, X_n$, and the random variables $X_1, X_2, \ldots, X_n$ are mutually independent. To test whether there is a trend in the data; that is, whether the later random variables are greater than the earlier random variables or

vice versa, the null hypothesis is: no trend is present. A one-sided test is used to detect an upward (or a downward) trend. The two-sided test is used if the alternative hypothesis is that any type of trend exists.

The random variables are grouped into pairs. If $n = 2m$ is even, then form $m$ groups as $(X_1, X_1 + m), (X_2, X_2 + m), \ldots, (X_m, X_{2m})$. If $n$ is odd, then discard the middle observation and use the remaining numbers to form groups in the same way. Next replace each pair $(X_i, X_{i+m})$ with a "+" if $X_i < X_{i+m}$ or a "−" if $X_i > X_{i+m}$, eliminating ties. Once these are done for all the pairs, one can follow the same procedure of the simple sign test given previously using the test statistic $T =$ total number of "+"s. Under the null hypothesis, $T \sim$ binomial $(n; 1/2)$. A preponderance of plus signs supports the hypothesis of an upward trend of $X_1, X_2, \ldots, X_n$, and a preponderance of minus signs suggests the presence of a downward trend.

### *Example*

The length of growth season is recorded for 21 years. The length in days is 224, 253, 216, 204, 231, 225, 236, 200, 243, 260, 215, 262, 233, 215, 242, 249, 227, 254, 225, 208, and 246. To determine whether these data provide sufficient evidence to indicate the presence of a trend in the length of growth season, hypothesize that there is no reason to suspect one type of trend (either upward or downward). As $n = 21$ is odd, the middle number 215 is omitted. Pairing the $i$th observation of the first half with the $i$th observation of the second half yields the following: (224, 262), (253, 233), (216, 215), (204, 242), (231, 249), (225, 227), (236, 254), (200, 225), (243, 208), and (260, 246). There are no ties, so $n = 10$ and $T = 6$. The $p$ value for the one-sided upper tail test is

$$\sum_{i=6}^{10} \binom{10}{i} \left(\frac{1}{2}\right)^{10} = 0.38,$$

which is nonsignificant. There is no evidence for the presence of a trend.

## Similar Tests

As a nonparametric test for paired observations, the sign test is used when the null hypothesis is that there are equal numbers of differences in each direction. There are other tests for paired observations of an ordinal or continuous variable. If the pair difference is a continuous variable and normally distributed, and the null hypothesis is that the mean of the difference between pairs of observations is zero, then the paired sample $t$ test can be used. If the null hypothesis is that the median of the difference between pairs of observations is zero, then the Wilcoxon signed rank test can be used.

*Hong Wang*

***See also*** McNemar's Test; Nonparametric Statistics; Repeated Measures Design; Significance, Statistical; Significance Level, Concept of; Student's $t$ Test; $t$ Test, One Sample; $t$ Test, Paired Samples

### Further Readings

Arbuthnott, J. (1710). An argument for divine providence taken from constant regularity observ'd in the births of both sexes. *Philosophical Transactions of the Royal Society of London, 27,* 186–190.

Cox, D. R., & Stuart, A. (1955). Some quick sign tests for trend in location and dispersion. *Biometrika, 42,* 80–85.

Hollander, M., & Wolfe, D. A. (1999). *Nonparametric statistical methods* (2nd ed.). New York: Wiley.

# Simple Main Effects

A great deal of psychological research examines the effect a single independent variable has on a single dependent variable. However, it is perfectly reasonable to examine the effects multiple independent variables simultaneously have on a single dependent variable. Experimental designs that include multiple independent variables are known as factorial designs. They come in a wide variety including between-group and within-subject variables and a mixture of both kinds. When two or more variables in a factorial design show a statistically significant interaction, it is common to analyze the simple main effects. Simple main effects analysis typically involves the examination of the effects of one independent variable at different levels of a second independent variable.

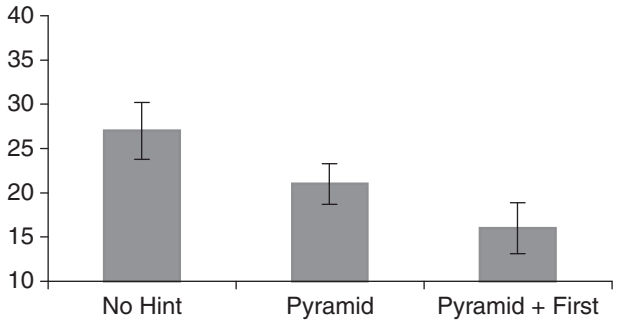



**Figure 1** The Means and Standard Errors for Different Hint Conditions for Solving the Tower of Hanoi Problem

## Interactions

Suppose an experiment was conducted that examined the consequences of providing different hints for solving the tower of Hanoi problem. The first condition simply provides the rules and the goal for the problem (*No hint*). The second condition involves telling the participants not to think about moving disks but moving pyramids of disks (*Pyramid*). The third condition involves telling participants about moving pyramids of disks and states what the first move is (*Pyramid + First*). The dependent variable is the number of moves to solution of the problem. The results are shown in Figure 1.

These results are explained in terms of reducing the amount of planning that is required. The greater the amount of planning that is necessary, the more mistakes that are made because it is easier to forget long plans.

A different experiment using the tower of Hanoi was also conducted. Three groups were selected according to age: 5–7 years, 9–11 years, and 13–15 years old. Figure 2 shows the results.

The results are also explained with respect to planning such that younger children cannot remember the plans they have made, whereas older children can. In the first experiment, the reduction in planning demands is obtained by changing the representation of the problem. In the second experiment, the ability to deal with planning demands vary according to age. To determine whether reducing the planning demands for younger children by changing the representation of the problem helps them to overcome their problems in planning, the two variables (age and problem representation) are considered at the same time.

Suppose the results shown in Figure 3 were obtained. In this case, the effects of changing the representation of the problem differ with respect to age. In other words, the effects of one variable are different for the different levels of the second variable. This is known as an interaction.

A more formal definition of an interaction is as follows:

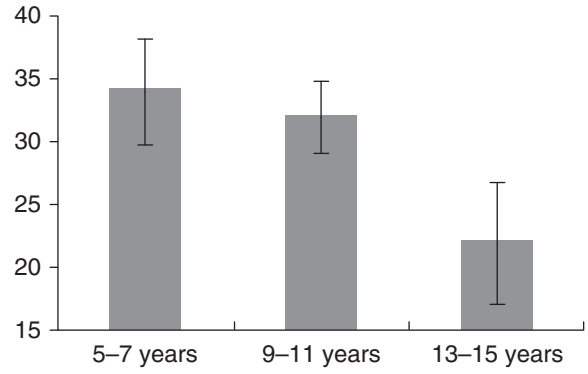



**Figure 2** The Mean and Standard Errors for Different Ages for Solving the Tower of Hanoi Problem

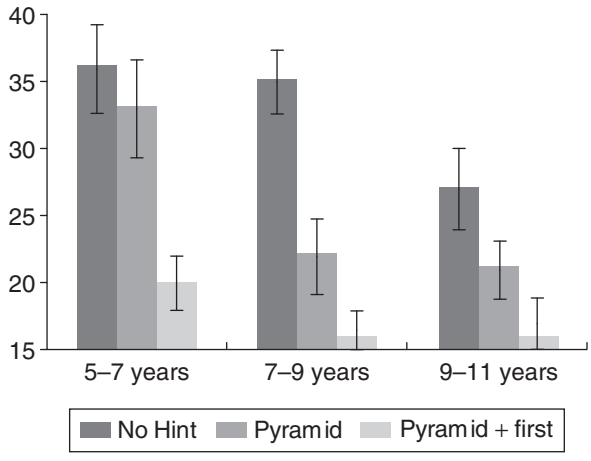



**Figure 3** The Mean and Standard Errors for the Interaction Between Different Ages and Hint Conditions for Solving the Tower of Hanoi Problem

**Table 1** An Example ANOVA Output

| *Source* | *Sum of Squares* | df | *Mean Square* | F *ratio* | p *value* |
|---|---|---|---|---|---|
| Learning | 3.20 | 1 | 3.20 | 2.56 | .13 |
| Task | 105.80 | 1 | 105.80 | 84.64 | $8.65 \times 10^{-8}$ |
| Learning * Task | 7.20 | 1 | 7.20 | 5.76 | .03 |
| Error | 20.00 | 16 | 1.25 | | |
| Total | 136.20 | 19 | | | |

An interaction is present when the simple effects of one independent variable are not the same at different levels of the second independent variable.

The simple effects of an independent variable are the effects of one independent variable at the specific levels of the other independent variable.

## Analyzing Simple Main Effects

When an analysis of variance (ANOVA) is used to examine interactions, both independent variables are entered into the analysis simultaneously. The output from most statistical packages resembles that shown in Table 1.

The following is an example of a between groups analysis. In this case, the first independent variable is expressed as two groups of participants. The first is presented with instructions that lead to learning an artificial grammar implicitly (*implicit condition*), whereas the second is given instruction to learn an artificial grammar explicitly (*explicit condition*). The second independent variable involves one group of participants working on their task with no further task (*single task condition*) with a second group given a secondary task to complete at the same time (*dual task condition*). The dependent variable is the number of correctly identified, novel grammatical items at posttest. Altogether, there are four separate conditions. The results are shown in Figure 4.

In this analysis, there is no evidence that the learning manipulation has led to a difference in the implicit and explicit condition means for the number of correctly identified grammatical items. At the same time, there is a significant main effect of task. That is, the mean number of correctly identified grammatical items is less in the single task condition than in dual task condition. Finally, there is a significant interaction between the task and learning variables. By simply looking at the graphs, it is not immediately obvious how this interaction works. Instead, we need to examine statistically the interaction even more to understand how the means differ.

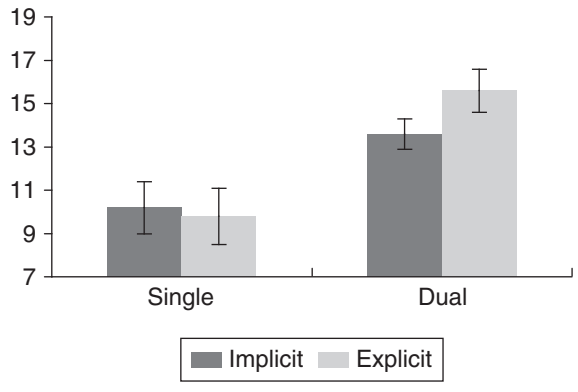



**Figure 4** The Means and Standard Errors for the Interaction Between Learning and Task Instructions

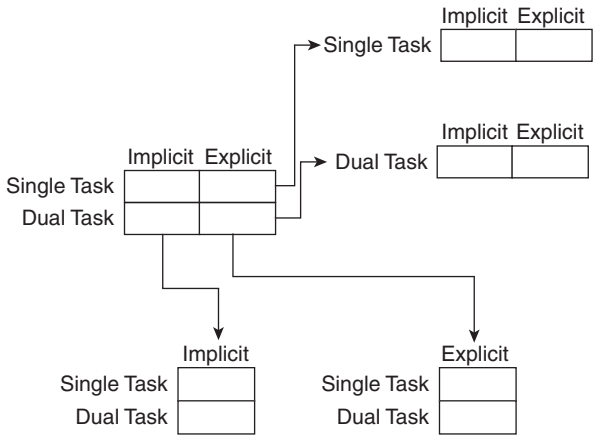



**Figure 5** A Breakdown of a Two-Way Interaction Into Its Constituent Simple Effects

**Table 2** An Example Simple Effects Analysis

| *Source* | *Sum of Squares* | df | *Mean Square* | F *ratio* | p *value* |
|---|---|---|---|---|---|
| Learning at | | | | | |
| Single | 0.40 | 1 | 0.40 | 0.32 | .58 |
| Dual | 10.00 | 1 | 10.00 | 8.00 | .01 |
| Task at | | | | | |
| Implicit | 28.90 | 1 | 28.90 | 23.12 | $1.93 \times 10^{-4}$ |
| Explicit | 84.10 | 1 | 84.10 | 67.28 | $4.00 \times 10^{-7}$ |
| Error | 20.00 | 16 | 1.25 | | |

The most commonly used technique for the subsequent analysis of interaction is known as simple effects analysis. It is sometimes also known as simple main-effects analysis. The simplest way to understand simple effects analysis is to consider the overall experiment being broken down into several smaller constituent experiments. For example, the example given previously can be broken down into four constituent experiments (see Figure 5). First, the simple effect of learning at single task can be considered. This involves comparing the implicit and explicit condition means for the single task condition alone. Then, there is the simple effect of learning at dual task. The means for the implicit and explicit conditions are contrasted for just the dual task data. In contrast, there is the simple effect task at implicit learning examines the difference in the means for the single task and dual task means for the implicit data by itself. Finally, the simple effect of task at explicit learning compares the means for the explicit condition between the single and dual task conditions. Typically, simple effects are referred to as the effect of variable A at levels 1 to *k* of variable B and vice versa.

For each of the simple effects that represent the interaction, a separate *F* ratio is calculated. A typical completed analysis looks like Table 2.

The simple effect of learning at a single task is not significant. There is, therefore, no evidence to support the hypothesis that implicit and explicit learning condition means are different when a single task is presented. However, the remaining three simple effects are statistically significant. This implies that the effect of a dual task is to increase the means when compared with a single task and that this increase is sufficiently large for it to lead to a difference between the means for the dual task condition between the implicit and explicit learning conditions.

## Simple Effects Analysis for More Complex Designs

The previous example is relatively straightforward in the sense that there were only two independent variables, and each of those independent variables only had two levels (or conditions). The situation is more complicated when at least one of the independent variables has more than two conditions. When this happens, the *F* ratio associated with the simple effect with more than two conditions tests whether overall those three or more means are significantly different from each other. It does not test directly how those means actually differ. In this case, two basic approaches are often taken. First, if the pattern of mean differences had been predicted prior to the experiment, it is possible to use planned (a priori) contrasts to test the differences between the means. If, however, the mean differences had not been predicted prior to analysis, it is common to use unplanned (post hoc) comparisons. In both cases, it is important not to conduct unnecessary statistical tests because this rapidly increases the chance of making a Type I error. Planned comparisons are usually conducted using the method of weighted means and are sometimes referred to as *contrasts*. Post hoc comparisons are somewhat more problematic. There is no overall agreement in the literature about which post hoc test is best used for which experimental design. However, the Tukey's Honestly Significant Difference test has been adapted to work fully between groups designs, fully within subjects designs, and designs in which there is both a between-group and within-subject component.

A second way in which designs can be more complicated is when there are more than two independent variables. In this case, not only might two factor interactions be found but also higher level

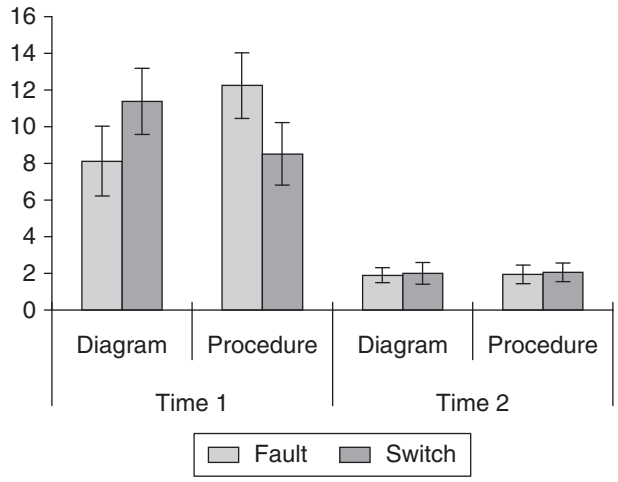



**Figure 6** The Means and Standard Errors for the Three-Way Interaction Among Instruction, Task, and Time

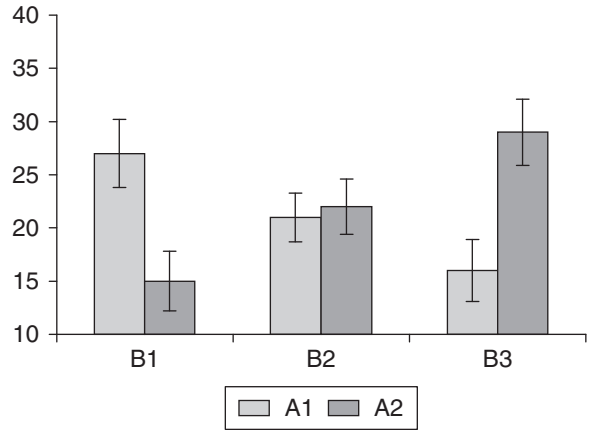



**Figure 7** The Means and Standard Errors for the Interaction Between Two Trends

interactions between three or more factors. Fortunately, the analytic strategy that is adopted is simply an extension of what has already been outlined.

For example, Figure 6 shows a two-way interaction between instruction type (diagram vs. procedures) and task type (fault vs. switch) at time 1. However, that interaction is no longer obvious at time 2. These data show a three-way interaction. One way to understand such an interaction is that the subsumed two-way interactions are not the same. An analytic strategy that examines the two-way interactions, therefore, can help test this hypothesis. The simple effects of task at (diagram, time 1), task at (procedure, time 1), instruction at (fault, time 1) and instruction at (switch, time 1) can be examined as well as the equivalent simple effects at time 2. In this way both, the two-way interactions are examined and can provide a helpful insight into the three-way interaction.

## Choosing the Appropriate Error Term for Simple-Effects Analysis

One important concern when conducting a simple-effects analysis is choosing the appropriate error term. For the standard between-groups analysis, the error used when calculating the *F* ratio for the simple effects is usually the same error used in the overall ANOVA. This is the pooled error term calculated by combining all the data in the experiment. However, when there are heterogeneous variances across the levels of an independent variable, comparisons are made such that the actual variability when comparing conditions is either significantly greater or less than the average variability. When this happens, the observed *F* ratio can be biased and lead to either Type I or Type II errors. It is usually safer in this case to use an error term that is based solely on the variability around the means that are actually being compared. In a two-way between-groups design this is equivalent to extracting the appropriate data from the design and conducting the equivalent of an independent samples *t* test for two levels or a one-way between groups ANOVA when there are more than two levels.

In a completely within-subjects design, the overall error term is used as is normally the case in the completely between-groups design. However, for a mixed design with both between-groups and within-subjects variables, the error that is adopted depends on whether a between-groups simple effect or a within-subject simple effect is being examined. When examining the between-groups simple effect, a more accurate estimate of the error variability is obtained using a pooled within-cell error term, which has more degrees of freedom. This assumes that there is homogeneity of within condition variances. This error term is often larger

than the usual simple effect error term, but the critical value of the *F* ratio will be smaller given the greater degrees of freedom, thus giving more statistical power to the test. If there is heterogeneity of within-condition variances, then this is not appropriate and separate error terms are calculated for each level of the within-subject variable.

### Alternatives to Simple Effects Analysis

One problem with using simple effects analysis, particularly with complicated designs, is the proliferation of statistical tests and the ensuing increase in familywise Type I error rates. As more tests are conducted, the chance of falsely finding a significant result also increases. An alternative to this strategy is to conduct what are known as interaction contrasts. Figure 7 shows an example where such a method is particularly useful.

In this case, there is a descending linear trend for level A1 and an increasing linear trend for level A2 across the three B treatment levels. In this case, an interaction contrast can be constructed that directly tests the hypothesis that the two trends are statistically different.

*Peter Bibby*

***See also*** Analysis of Variance (ANOVA); A Priori Tests; Post Hoc Analysis; Post Hoc Comparisons; Tukey's Honestly Significant Difference (HSD); Type I Error; Type II Error

#### Further Readings

Keppel, G., & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook* (4th ed.). Upper Saddle River, NJ: Pearson Education Inc.

Kirk, R. E. (1995). *Experimental design: Procedures for behavioral sciences* (3rd ed.). Belmont, CA: Wadsworth.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective*. Mahwah, NJ: Lawrence Erlbaum.

## Simpson's Paradox

The marginal association between two categorical variables might be qualitatively different than the partial association between the same two variables after controlling for one or more other variables. This is known as Simpson's paradox. Simpson's paradox is important for three critical reasons. First, people often expect statistical relationships to be immutable. They often are not. The relationship between two variables might increase, decrease, or even change direction depending on the set of variables being controlled. Second, Simpson's paradox is not simply an obscure phenomenon of interest only to a small group of statisticians. Simpson's paradox is actually one of a large class of association paradoxes. Third, Simpson's paradox reminds researchers that causal inferences, particularly in nonexperimental studies, can be hazardous. Uncontrolled and even unobserved variables that would eliminate or reverse the association observed between two variables might exist.

### Illustration

Understanding Simpson's paradox is easiest in the context of a simple example. The inspiration for the one presented here is the classic article published by Peter Bickel and his colleagues.

Suppose that a university is concerned about gender bias during the admission process to graduate school. To study this, applicants to the university's graduate programs are classified based on gender and admissions outcome. These data, which are summarized in the cross-classification table depicted in Table 1, would seem to be consistent with the existence of a gender bias because men (40%) were more likely to be admitted to graduate school than women (25%).

To identify the source of the difference in admission rates for men and women, the university subdivides applicants based on whether they applied to a department in the natural sciences or to one in the social sciences and then conducts the analysis again (see Table 2). Surprisingly, the

**Table 1** Graduate Admissions by Gender

| | *Admit* | *Deny* | *Total* |
|---|---|---|---|
| Women | 31 | 95 | 126 |
| Men | 66 | 100 | 166 |
| Total | 97 | 195 | 292 |

**Table 2** Graduate Admissions by Gender by Department

| *Social Science* | | | | *Natural Science* | | | |
|---|---|---|---|---|---|---|---|
| | *Admit* | *Deny* | Total | | *Admit* | *Deny* | Total |
| Women | 23 | 93 | 116 | Women | 8 | 2 | 10 |
| Men | 1 | 25 | 26 | Men | 65 | 75 | 140 |
| Total | 24 | 118 | 142 | Total | 73 | 77 | 150 |

university finds that the direction of the relationship between gender and outcome has reversed! In natural science departments, women (80%) were more likely to be admitted to graduate school than men (46%); similarly, in social science departments, women (20%) were more likely to be admitted to graduate school than men (4%).

Although the reversal in association that is observed in Simpson's paradox might seem bewildering, it is actually straightforward. In this example, it occurred because both gender and admissions were related to a third variable, namely, the department. First, women were more likely to apply to social science departments, whereas men were more likely to apply to natural science departments. Second, the acceptance rate in social science departments was much less than that in natural science departments. Because women were more likely than men to apply to programs with low acceptance rates, when department was ignored (i.e., when the data were collapsed over department), it seemed that women were less likely than men to be admitted to graduate school, whereas the reverse was actually true. Although hypothetical examples such as this one are simple to construct, numerous real-life examples can be found easily in the social science and statistics literatures.

## Definition

Consider three random variables $X$, $Y$, and $Z$. Define a $2 \times 2 \times K$ cross-classification table by assuming that $X$ and $Y$ can be coded either 0 or 1, and Z can be assigned values from 1 to $k$.

The *marginal association* between $X$ and $Y$ is assessed by collapsing across or aggregating over the levels of $Z$. Three possible patterns of association might be distinguished.

$$P(Y = 1|X = 1) > P(Y = 1|X = 0),$$
$$P(Y = 1|X = 1) = P(Y = 1|X = 0),$$
$$P(Y = 1|X = 1) < P(Y = 1|X = 0).$$

The *partial association* between $X$ and $Y$ controlling for $Z$ is the association between $X$ and $Y$ at each level of $Z$ or after adjusting for the levels of $Z$. Simpson's paradox is said to have occurred when the pattern of marginal association and the pattern of partial association differ. In the example discussed earlier, if women are coded 1 and men are coded 0, and admits are coded 1 and denies are coded 0, then the marginal association between gender and admissions will correspond to the third of the three possible patterns of association, whereas the partial association between gender and admissions for the two departments separately will correspond to the first pattern.

Various indices exist for assessing the association between two variables. For categorical variables, the odds ratio and the relative risk ratio are the two most common measures of association. Simpson's paradox is the name applied to differences in the association between two categorical variables, regardless of how that association is measured, whether descriptively in terms of conditional probabilities, as provided previously, or through one of the measures of association.

## Association Paradoxes

Association paradoxes, of which Simpson's paradox is a special case, can occur between continuous as well as categorical variables. For example, the best-known measure of association between two continuous variables is the correlation coefficient. It is well known that the marginal correlation between two variables can have one sign, whereas the partial correlation between the same two variables after controlling for one or more additional variables has the opposite sign.

*Reversal paradoxes,* in which the marginal and partial associations between two variables have different signs, such as Simpson's paradox, are the most dramatic of the association paradoxes. A weaker form of association paradox occurs when the marginal and partial associations have the same sign, but the magnitude of the marginal association falls outside of the range of values of the partial

associations computed at individual levels of the variable(s) being controlled. These have been termed *amalgamation* or *aggregation paradoxes*.

## Problem of Causality

When confronted with a reversal paradox, it is natural to ask whether the marginal or the partial association is the correct description of the relationship between two variables. Assuming that the relationships among the variables in one's sample mirror those of the population from which the sample was drawn, then the usual statistical answer is that both the marginal and partial associations are correct. Mathematically, there is nothing surprising about a reversal in the direction of the marginal and partial associations. Furthermore, in an analysis, such as the one presented previously, the reversal of the marginal and partial associations is easily understood once the role of the control variable is understood.

If social scientists were merely interested in cataloging the relationships that exist among the variables that they study, then the answer given previously might be sufficient. It is not. Often, social scientists are interested in understanding causal relationships. In the example given previously, one might be interested in knowing whether the admissions process is biased toward males, as the marginal association might suggest, or biased toward females, as the partial association might suggest. This is the real dilemma posed by Simpson's paradox for the researcher. It is problematic in two ways.

First, the statistical analysis provides no guidance as to whether the marginal association or the partial association is the spurious relationship. Based on our knowledge of graduate admissions, it is reasonable to conclude that the marginal relationship in this example is spurious because admissions decisions are made by departments, not by universities. Substantive information guides this judgment, not the statistical analysis. It might be tempting to conclude, as some authors do, that the marginal association is always spurious. Certainly, that is the impression that is given by much of the published work on Simpson's paradox. Indeed, some authors characterize Simpson's paradox as a failure to include a relevant covariate in the design of a study or in the relevant statistical analysis. Unfortunately, this simple answer is inadequate, because it is possible to construct examples in which the partial association is the spurious one.

Second, the field of statistics provides limited assistance in determining when Simpson's paradox will occur. Particularly in nonrandomized studies, there might exist uncontrolled and, even more dangerously, unobserved variables that would eliminate or reverse the association observed between two variables. It can be unsettling to imagine that what is believed to be a causal relationship between two variables is found not to exist or, even worse, is found to be opposite in direction once one discovers the proper variable to control.

## Avoiding Simpson's Paradox

Although it might be easy to explain why Simpson's paradox occurs when presented with an example, it is more challenging to determine when Simpson's paradox will occur. In experimental research, in which individuals are randomly assigned to treatment conditions, Simpson's paradox should not occur, no matter what additional variables are included in the analysis. This assumes, of course, that the randomization is effective and that assignment to treatment condition is independent of possible covariates. If so, regardless of whether these covariates are related to the outcome, Simpson's paradox cannot occur. In nonexperimental, or nonrandomized, research, such as a cross-sectional study in which a sample is selected and then the members of the sample are simultaneously classified with respect to all of the study variables, Simpson's paradox can be avoided if certain conditions are satisfied. The problem with nonexperimental research is that these conditions will rarely be known to be satisfied a priori.

## Brief History of Simpson's Paradox

Given the nature of the phenomenon, perhaps it is only fitting to discover that Edward Simpson neither discovered nor claimed to have discovered the phenomenon that now bears his name. In his classic 1951 paper, Simpson pointed out that association paradoxes were well known prior to the publication of his paper. Indeed, the existence of association paradoxes with categorical variables

was reported by George Udney Yule as early as 1903. It is for this reason that Simpson's paradox is sometimes known by other names, such as the Yule paradox or the Yule-Simpson paradox. It is possible to trace the existence of association paradoxes back even further in time to Karl Pearson, who in 1899 demonstrated that marginal and partial associations between continuous variables might differ giving rise to spurious correlations. Pearson reported that the length and breadth of male skulls from the Paris catacombs correlated .09. The same correlation among female skulls was −.04. After combining the two samples, the correlation was .20! In other words, skull length and breadth were uncorrelated for males and females separately and positively correlated for males and females jointly. Put slightly differently, the marginal association between skull length and breadth was positive, while the partial association between skull length and breadth after controlling for gender was zero.

Not only is Simpson not the discoverer of Simpson's paradox, but the phenomenon that he described in his 1951 paper is not quite the same as the phenomenon that is now known as Simpson's paradox. The difference is not critical, but it does reflect the confusion that persists today about what Simpson's paradox actually is. Some authors reserve the label Simpson's paradox for a reversal in the *direction* of the marginal and partial association between two categorical variables. Some authors apply Simpson's paradox to reversals that occur with continuous as well as categorical variables. Still other authors have abandoned the term *Simpson's paradox* altogether, preferring terms such as *aggregation, amalgamation,* or *reversal paradoxes,* which are often defined more broadly than Simpson's paradox.

*Bruce W. Carlson*

***See also*** Association, Measures of; Categorical Variable; Confounding; Ecological Validity; Odds Ratio; Partial Correlation; Regression Artifacts

**Further Readings**

Appleton, D. R., French, J. M., & Vanderpump, M. P. J. (1996). Ignoring a covariate: An example of Simpson's paradox. *American Statistician, 50,* 340–341.

Bickel, P. J., Hammel, E. A., & O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science, 187,* 398–404.

Mantel, N. (1982). Simpson's paradox in reverse. *American Statistician, 36,* 395.

Paik, M. (1985). A graphic representation of a three-way contingency table: Simpson's paradox and correlation. *American Statistician, 39,* 53–54.

Samuels, M. (1993). Simpson's paradox and related phenomena. *Journal of the American Statistical Association, 88,* 81–88.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B,* 13, 238–241.

Wainer, H., & Brown, L. M. (2004). Two statistical paradoxes in the interpretation of group differences: Illustrated with medical school admission and licensing data. *American Statistician, 204,* 58, 117–123.

# Single-Blind Study

A single-blind study occurs when the participants are deliberately kept ignorant of either the group to which they have been assigned or key information about the materials they are assessing, but the experimenter is in possession of this knowledge. Single-blind studies are typically conducted when the participants' knowledge of their group membership or the identity of the materials they are assessing might *bias* the results. However, there are situations where creating such ignorance might be impossible or unethical, and in others, it might be advisable for more than the participants to be kept unaware of the test conditions. This entry discusses the single-blind study in relation to the unblinded (neither experimenter nor participants are kept ignorant) and double-blind study (both experimenter and participants are kept ignorant). It also presents some illustrative examples and examines the advantages and disadvantages of the single-blind study.

## The Unblinded Study

To understand the blind study, it is advisable first to consider the unblinded study. In this, both participant and observer are fully aware of the groupings and/or materials. For example, a group of

patients might be fully aware of which of two types of treatment they are receiving, such as in judging a group of singers in a competition, the participants are fully aware of the identity of each singer. Sometimes, an unblinded study is unavoidable. For example, many routine clinical treatments are done in an unblinded manner, with the patient and therapist actively discussing all the methods available before choosing one. Accordingly, any naturalistic, purely observational study of such a situation will of necessity have to accept that the participant is aware of what is taking place. Likewise, in some circumstances, asking participants to make judgments will also require them to know the identity of what is being judged (e.g., judging the relative merits of famous singers). The unblinded study is, thus, in some instances unavoidable. In practical terms, it is also generally far easier to run than a blind study, which requires greater organization.

However, the unblinded study also carries disadvantages that sometimes outweigh its merits, and experimenters will usually avoid it when a blind study is a viable alternative. At its most basic, the unblinded study provides the participant with information about how they are grouped or what they are assessing, and this knowledge might lead to self-expectations of how they should respond. This can potentially distort results, as is discussed next.

## Illustrative Examples

In contrast to unblinded studies, single-blind studies deliberately withhold key information from the participants (though not the experimenters). A commonplace example of withholding key information about test materials is a measure of consumer preference where the participant is asked to judge which of two or more brands of the same basic product is preferred, but the brand identities of the products are hidden from the participant. For example, in comparing the quality of different hi-fi components, products could be hidden behind visually opaque but acoustically transparent material, and the participants could be asked to listen to music played over these systems without being aware of the brands of product. All the participant can do is judge the quality of what they hear. However, the experimenters are aware of the true identity of each piece of equipment being assessed. An example of participants lacking knowledge of which group they belong to is the common clinical procedure where patients are given one of two or more drug treatments, but they do not know which one. However, the experimenters providing the treatments and assessing the treatment are aware of this information.

## Advantages

The key advantage of the single-blind study is that, all things being equal, the participant is deprived of the biasing effect that knowledge of the test materials or group membership might bring. For example, in the case of the hi-fi experiment, because the participant does not know the brand of the products being tested, it is assumed that the participant cannot be swayed in his or her judgment by brand loyalty or brand status. Thus, if expensive and inexpensive pieces of hi-fi equipment are being compared, the former cannot be automatically preferred because it is more expensive and hence "must" be better, because the participant cannot see which product is which and hence is compelled to judge on sound alone, rather than sound distorted by biased expectations. In the drug trial experiment, the knowledge of which treatment a person is receiving could have a biasing effect. For example, suppose that the patient knows that he or she is being given an established (and not very effective) drug instead of the new drug that the other group is receiving. The patient might feel depressed that he or she is getting a less effective treatment, and this depression might lessen the patient's recovery rate. Conversely, if a patient feels suspicious of new methods and prefers the tried and tested, then receiving the old drug might make him or her feel more optimistic and might artificially elevate his or her mood and, hence, aid recovery. These are examples of what in loose terminology are called *nocebo* and *placebo* responses (lowering or raising of beneficial effects through expectation alone). The placebo and related effects are central to the single-blind study, and the dangers (or perceived dangers) of bias through participant expectation lie at the heart of many blind study designs. By removing knowledge of group assignment, expectation is also removed and, hence, it is argued, possible biasing effects.

## Disadvantages

However, nocebo and placebo responses cannot be quite so easily disposed of. It is possible that *any* treatment a person receives might have an effect, even if it is supposed to be totally inert. For example, a drug might have part of its effect, not because of its pharmacological function, but because the patient expects the drug to have an effect and thus will change the patient's mood accordingly. Typically, patients expect to get better after treatment, and so the simple effect of taking their medicine regularly might cause an improvement in reported mood and health because they expect to be feeling better. To counteract this, single-blind studies might contain a placebo group. In other words, a group whose treatment is inert (literally, in the case of a drug trial—the placebo "drugs" taken have no pharmacological effect). This enables experimenters to judge how much of the effect of treatment is a result of the simple act of administering a treatment. For one of the real treatments to be effective, it has to offer advantages above and beyond that conferred by the placebo dose. However, there are ethical limitations to placebo treatments. If patients have a life-threatening condition, for example, it is morally very difficult to deny a patient an active treatment and offer them a sham therapy instead. A related issue is whether patients are treated equally in all other respects. The argument has been raised that if a clinician believes that a patient is receiving an inferior treatment to that being given to patients in another group, he or she might be tempted (for entirely understandable reasons) to give the patient other therapies and drugs to compensate. This of course will distort the results, because it will artificially diminish the real size of the difference between the two drugs. This might bring benefits to the patients in the study, but if the better drug is not then adopted because it (spuriously) seems to be not sufficiently better than other available treatments, then the entire population of patients suffers.

Related to this point, another key problem with the single-blind study is that although the participants might be naïve about what they are assessing or the groups into which they have been placed, the experimenters are not. The key problem is *experimenter bias* (a type of *experimenter effect*). In other words, the experimenters might unwittingly bias the results either through influencing the participants or in recording their own assessments. To return to the drug trial example: suppose that the drugs being tested are antidepressants, and the experimenters have the task of assessing the patients' level of depression during the trial. If the experimenters have a vested interest in the outcome of the experiment, they might unwittingly be drawn toward observing bigger improvements in those patients who receive the drug therapy they most support. In the hi-fi equipment example, if the experimenters have a vested interest in one of the products, then they might, through body language or suggestion, imply that "their" product is better than the others. In addition, factors such as the sequencing of presentation might also have a biasing effect. Furthermore, there are potentially cases of outright fraud or of less-than-honest practices (e.g., comparing two brands of soda pop, and serving the one intended to win at a cooler temperature). Activities akin to this have been rumored to take place in *some* less-than-scrupulous consumer trials intended to produce spurious statistics to support a marketing campaign. However, arguably this is a problem of the probity of experimenters and not the single-blind study per se.

## The Double-Blind Procedure

The single-blind study is potentially impervious to these criticisms if, issues of experimenter honesty aside, the measures being taken are entirely objective (e.g., subjective opinions of a particular drug's efficacy are unlikely to matter if the drug is a growth hormone and all measurements are objective ones of length). However, such situations rarely arise and critics can nearly always find real or potential sources of experimenter bias. Therefore, it is generally desirable to remove experimenter bias as far as possible by using a double-blind procedure. In essence, the double-blind method ensures that not only the participants but also the experimenters do not know to which group the participants belong. Thus, any observations by the experimenters must by definition be objective. As they do not

know to which group a participant belongs, there is no impetus to distort the findings.

However, although the double-blind procedure is generally presented by authors as being automatically superior to the single-blind study, there are some caveats to consider. The first is that just as the single-blind study is more complex to run than an unblinded study, so in turn is the double-blind study more difficult to run than a single-blind study. If the potential for experimenter bias justifies this, then it should be done, but if the measures are arguably objective enough to withstand any biasing effects, then running a double-blind measure seems a little excessive and unnecessarily wasteful of time and resources. It should also be noted that, in double-blind studies of clinical trials, a lack of knowledge about the treatment group to which a patient belongs could in certain circumstances be dangerous. Although unlikely, it is possible that a certain set of symptoms might give early warning of a serious side effect that the experimenter might notice if he or she knew the drug being administered, but which are merely noted as a set of minor symptoms in a blind observation. Again, in some cases, experimenter knowledge might be unavoidable. For example, if the observations to be made can only be made during the therapeutic process itself (and not, for example, simply in tests conducted after a drug has been administered), then the experimenter of necessity must be aware of the treatment he or she is administering. It should also be noted that in some instances, even if the identity of the treatment is hidden from the experimenter, the experimenter might quickly identify it by other characteristics (e.g., side effects reported by patients might enable a clinical observer to make an accurate guess about the nature of the drug treatment).

*Ian Stuart-Hamilton*

***See also*** Bias; Double-Blind Procedure; Experimenter Expectancy Effect; Nonclassical Experimenter Effects; Placebo; Placebo Effect

**Further Readings**

Friedman, L. M., Furberg, C., & DeMets, D. L. (1998). *Fundamentals of clinical trials* (3rd ed.). New York: Springer.

## Single-Subject Design

A single-subject design, which is also known as a single-case research design, provides a rigorous approach for documenting experimental effects. Single-case research has been used to (a) establish basic principles of behavior, (b) document the impact of specific interventions, and more recently (c) establish evidence-based practices. The defining feature of single-case research is the use of each participant (subject) as his or her own experimental control. This approach to research design arose from early work of researchers such as Burrhus F. Skinner and Werner Honig in the 1940s and 1950s focusing on behavior analysis, and it was codified in the seminal book *Tactics of Scientific Research* by Murray Sidman in 1960. Sidman defined in detail how the systematic study of individual participants over time could be used to test important experimental concepts and expand our fundamental understanding of human behavior.

Single-case methods have become a standard approach for conducting educational, behavioral, and psychological scholarship. Although single-case research is associated with behavioral psychology (e.g., *Journal of Applied Behavior Analysis*), single-case methods are now used to document advances across an array of fields including social-learning theory, medicine, social psychology, education, social work, pediatric psychology, and communication disorders.

The basic goal of single-case research designs is to evaluate the extent to which a causal (or functional) relation exists between introduction of an "intervention" (e.g., pace of instruction) and change in a specific dependent variable (e.g., reading performance). In most cases, a researcher conducting a single-case study wishes to establish that (a) prior to intervention a subject behaved in a consistent, clear manner that does *not* meet social expectations (e.g., the student read slowly or inaccurately, or the student engaged in high rates of problem behavior); (b) after the intervention the participant behaved in a manner that is both different and better than preintervention (e.g., the student read faster and/or more accurately, or the student behaved with less problem behavior and more positive behavior); and (c) this clearly measured change is unlikely to be the result of

anything other than the intervention (e.g., not normal development, not some other unmeasured intervention, and not illness or change in the personal life of the participant). It is this final assertion that the change in behavior was *causally* associated with introduction of the intervention that establishes the scientific credibility of single-case designs. Here, the defining features of single-case experimental research designs are reviewed, noting their contribution to a science of human behavior.

## The Individual as Unit of Analysis

Single-case research involves the fine-grained analysis of change across time. The individual subject serves as the unit of analysis, and each individual is observed multiple times before and after intervention. An experimental effect is demonstrated through the consistency and replicability of documented effects with individuals. The different types of single-case designs are constructed around varying options for building and replicating this comparison of preintervention and postintervention performance of individuals.

Although the individual always serves as the core unit of analysis in a single-case design, the following two caveats are worth noting: (1) an individual might be a person or even a group (e.g., whole classroom of students or whole school) and (2) to demonstrate a convincing effect, it is expected that change will be replicated across multiple individuals. The replication of effect can occur in many ways, but in most cases it involves demonstrating that an effect observed with one individual is also observed with others. The importance of replication can be observed by examining the number of participants reported in single-case research articles published in Volume 40 (2007–2008) of the *Journal of Applied Behavior Analysis*. From a total of 33 studies applying single-case methods, the average number of participants per study was 5.12 (median = 3). The unit of analysis in single-case research is the individual participant, but confidence in the validity of effects with individuals is enhanced through replication with additional participants.

## Replicable Description of Participants and Setting

A core feature of any experimental research is a precise and replicable description of the research procedures. In addition to including a replicable description of experimental methodology, single-case research aides replicability through the detailed description of (a) the individuals involved in the research (age, disability, and how selected [e.g., based on level of problem behavior]) and (b) the context or setting where the research was conducted.

## Measurement of Dependent and Independent Variables

An underemphasized feature of single-case research is the rigorous measurement of both the dependent (the outcome measure) and independent (intervention) variables. Because single-case research focuses on the change of individual behavior across time, the designs require collecting data about the dependent variable at multiple points in time. This requires careful and repeated measurement of the dependent variable (e.g., reading performance or problem behavior). However, these measurements must occur with a level of precision and validity that allows the reader to trust the accuracy of the numbers. As such, single-case research results typically include not only replicable descriptions of how data were collected (and by whom) but also documentation that the data meet minimal standards for reliability (e.g., interobserver agreement).

Interest in treatment fidelity within the broader educational and social sciences research community has led to an expectation that formal research must document the fidelity with which the independent variable (the intervention) was implemented. Not only is it essential to define the procedures within an intervention using a level of precision that allows replication, but also it is important to document that the intervention was actually implemented as described. Only by confirming that an intervention was conducted with fidelity can the results (whether positive or negative) be accurately attributed to the independent variable. Measuring independent variable fidelity is not yet a "necessary" feature of single-case

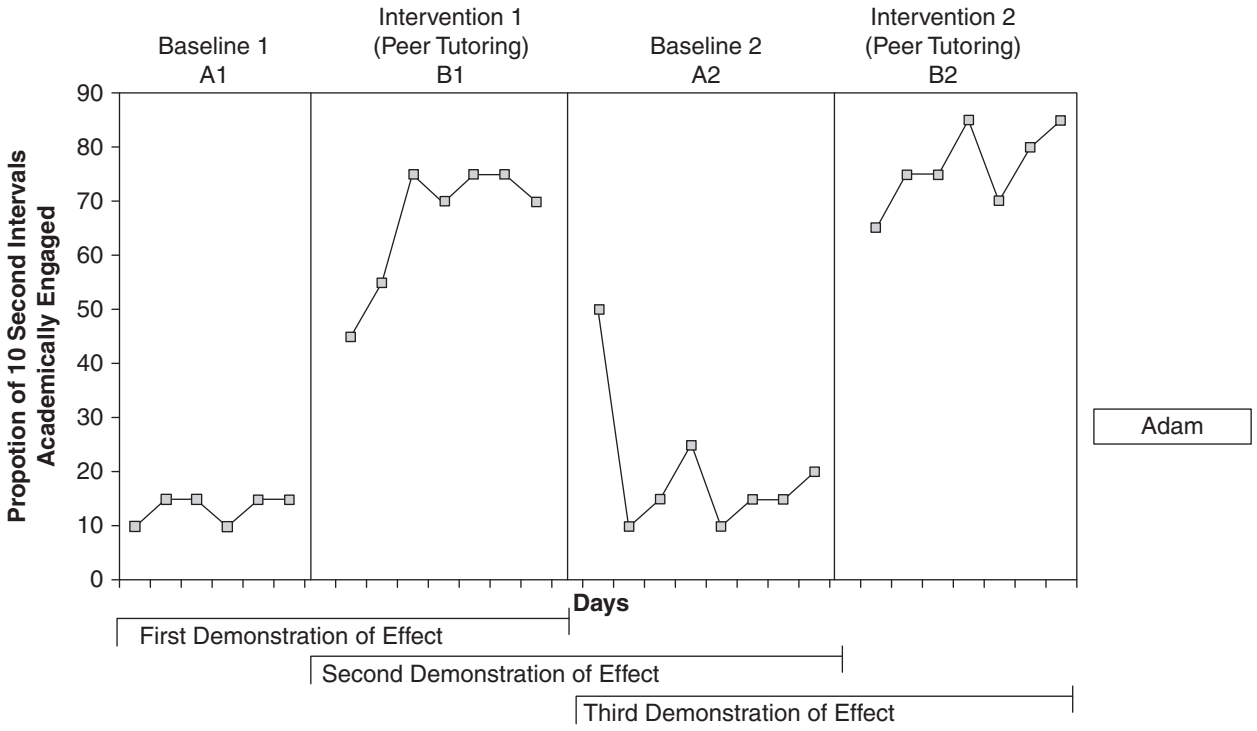



**Figure 1** Simulated Example of a Reversal Design Examining the Effects of Peer Tutoring on the Proportion of Time Adam Spends Academically Engaged

research, but it is rapidly gaining status as an expectation for publication.

## Baseline Documentation

Single-case research studies typically begin with a formal "baseline" period in which data are collected under "typical" or control conditions. The purpose of the baseline is to document a pattern of performance that (a) is different than is desired (too much problem behavior or too little learning) and (b) allows the prediction of future performance if no intervention were to occur.

The baseline condition is of tremendous value because it establishes the standard for comparison of the dependent variable across time. Documenting an adequate baseline pattern typically involves at least five sessions of baseline data (five data points), although more data points might be needed if high variability or trend in the data prevent identification of a stable pattern. Once baseline performance is established the experimenter manipulates (changes) only the independent variable (the intervention). This allows the researcher to ask: "If all other variables are held constant and the intervention is introduced, is there change in the dependent variable?" Documenting both the procedures of a baseline phase and a predictable pattern of behavior during baseline are core foundations of single-case research.

## Experimental Designs That Document Causal Relations

Experimental research is expected to (a) describe a phenomenon (e.g., behavior) with precision and clarity, (b) demonstrate that important and valued change has occurred with respect to that phenomenon, and (c) provide experimental controls allowing inference that the change was causally linked to the intervention. With single-case research, these goals are achieved through a variety of experimental designs, each of which allows *at least three demonstrations of effect across at least three different points in time*. Figures 1–4 provide examples of the four most common single-case

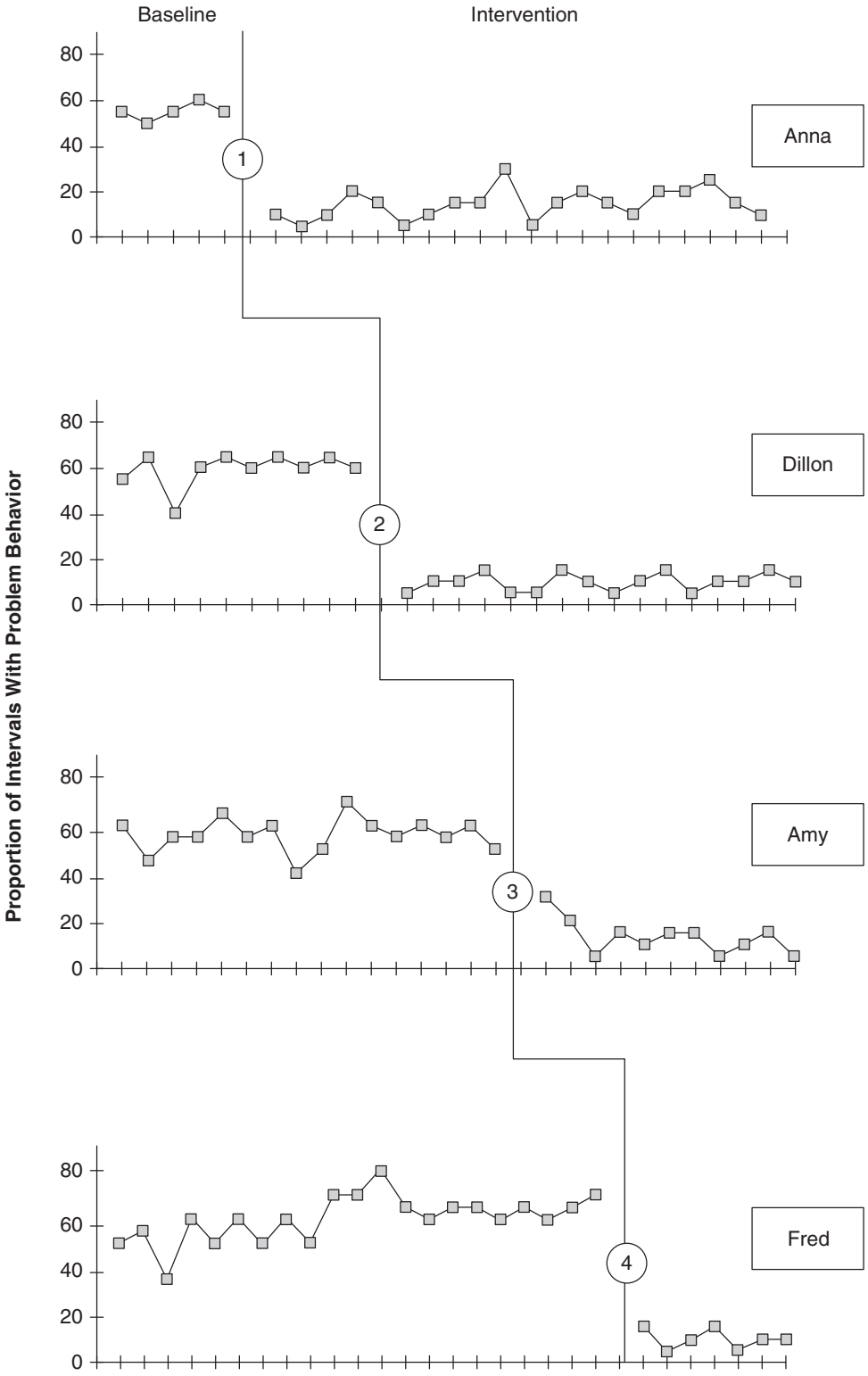



**Figure 2** Simulated Multiple-Baseline Design Across Students

*Note:* The design provides four demonstrations of clinical effects at four different points in time (indicated by numbers in circles).

designs, as well as include an index of how each demonstrates at least three demonstrations of effect across at least three different points in time.

### *Reversal or Withdrawal Designs*

The most commonly used single-case design is the reversal/withdrawal design shown in Figure 1.

As with most single-case designs, the dependent variable (outcome variable) is indexed on the ordinate (vertical axis) and time is indexed on the abscissa (horizontal axis). The design begins with documentation of baseline performance (A1). The first demonstration that manipulation of the independent variable is associated with an effect on the dependent variable occurs when the intervention is introduced (B1). The second demonstration of effect occurs with a return to the baseline condition (A2). If the dependent variable returns to baseline levels, there is increased confidence that the change in the dependent variable was causally related to the introduction of the intervention and not some other factors. The third demonstration of effect is provided when the intervention is reintroduced (B2) and the dependent variable again improves. By convention, three demonstrations of the effect at three different points in time is considered sufficient to infer that change in the dependent variable is functionally (or causally) related to the intervention. Note that in Figure 1, the brackets indicate the three demonstrations of effect at three different points in time.

It is important to note that a traditional case study in which data were collected in baseline (A1) and compared with performance during intervention (B1) would *not* qualify as an experimental single-case design. An A1-B1, or even an A1-B1-A2, design would provide a description of a problem and a documentation of change, but it would not offer three demonstrations of the effect at three different points in time. As such, these designs would not have the level of experimental control needed to demonstrate a causal or functional relation.

### *Multiple-Baseline Designs*

The multiple-baseline design addresses a major weakness of the reversal/withdrawal design: There are some dependent variables that cannot be reversed once the intervention has been successful (e.g., reading skills, steps completed correctly in a task analysis) or is clinically undesirable to reverse (e.g., severe self-injurious behavior). As the name implies, the multiple-baseline design is organized around multiple baselines (at least three and often more), with the intervention introduced at different points in time after each baseline has been established. Figure 2 provides an example of a four-series, multiple-baseline design.

Note that each baseline documents (a) a pattern of behavior that is undesirable and (b) sufficient consistency of responding (minimum of five data points) to allow prediction of responding into the future. The same intervention is introduced with each baseline series, but the point of intervention introduction is distributed across four different points in time. The introduction of the intervention in later series is delayed until there is a clear "effect" in the prior series. Taken together, the four baselines in Figure 2 meet the criterion of documenting at least three demonstrations of effect across at least three different points in time.

Multiple-baseline designs are often constructed around multiple participants, with the intervention introduced in a staggered timeline across participants. It is also possible, however, to organize multiple baseline designs with a single participant and three to four baselines constructed around different settings or activities.

### *Alternating Treatment/Multielement Designs*

A novel application of single-case logic is found in the alternating treatment (or multielement) design exemplified in Figure 3. The research questions addressed by alternating treatment/multi-element designs focus on comparing the effects of two or more "treatments" on a dependent variable. The design requires rapid alternating among treatments in a non-predictable ordering of presentation. If responding is markedly, and consistently, different in one treatment over others, then an inference is made that the difference in behavior is functionally related to the features of that treatment.

### *Changing-Criterion Design*

Changing-criterion designs are used less frequently and are more idiosyncratic in focus than the three designs noted previously. The research question addressed by a changing-criterion design is

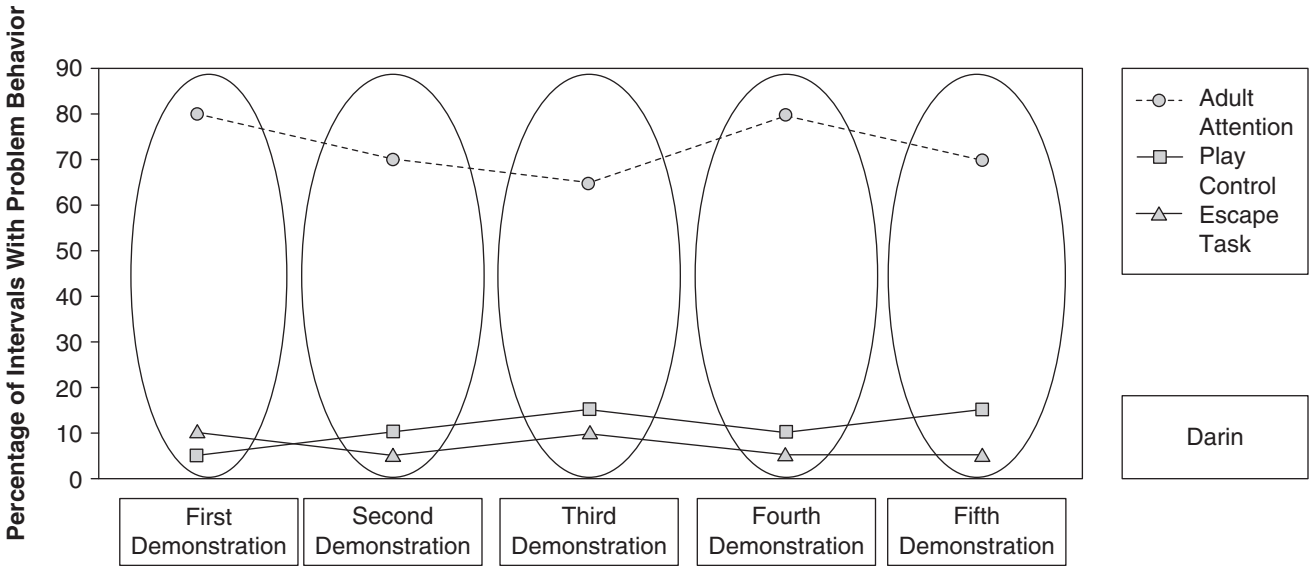



**Figure 3** Simulated Alternating Treatment Design Assessing the Effects of Different Consequences on Darin's Problem Behavior

whether the specification of a criterion, or limit (e.g., number of cigarettes smoked, number of math problems completed, or distance run), affects the targeted dependent variable. The focus of the analysis is on the specification of "a criterion" not the specific criterion. For example, if, as in Figure 4, a special reward was provided for reading two books in a month, and a child moved from 0 books per month to 2 books per month, then it could be argued that specification of the criterion (and reward) was related to the increase in books read. But because many other variables could have influenced the observed increase in reading, the effect must be replicated at least three times across at least three points in time. This could be done by increasing the criterion for reward from 2 to 3 books per month, and then from 3 to 4 books per month. If reading performance matched the criterion at *each* intervention criterion, then it is possible to infer that increased reading of books was functionally related to establishing a criterion for reward.

### *Combined Designs*

As research questions have become more complex, it is not uncommon for single-case researchers to ask multiple questions within one study and to incorporate more than one design model within a study. Figure 5 provides simulated intervention data in which multiple-baseline and reversal design elements are employed to evaluate first whether the intervention is effective, and second whether the intervention effects will maintain once the intervention features are removed. Ultimately, the design used in single-case research is driven by the research question (or questions) of interest. Although the basic design-logic of documenting three effects at three different points in time remains consistent across different design types, there are multiple strategies for achieving this standard.

## Visual Analysis

Single-case research relies on visual analysis of the data to determine whether a functional relation exists. The rules for conducting visual analysis are well documented but complex. The researcher must examine each phase of a single-case design by assessing the *level*, *trend*, and *variability* of the data within the phase. *Level* refers to the mean score for the data in a phase, *trend* refers to the slope of the best-fit straight line for the data within

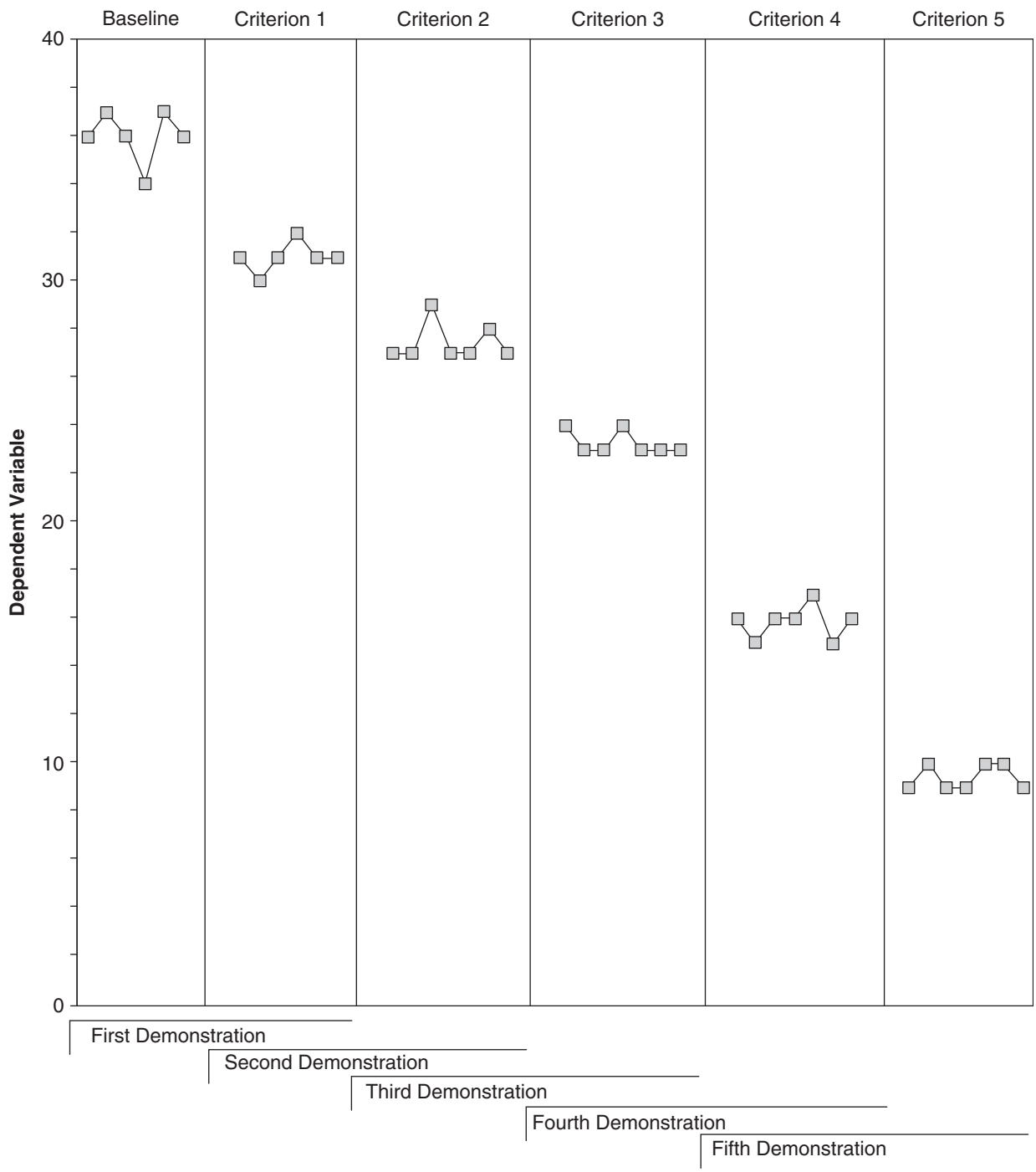



**Figure 4** Simulated Changing Criterion Design Providing Five Demonstrations of Effect at Five Different Points in Time

a phase, and *variability* refers to the range or standard deviation of data about the best-fit straight line. The examination of the data within a phase is used to describe both the observed pattern of performance during the observation sessions and to project expected performance if no changes were to occur. This assessment is then used to compare the observed and projected patterns with

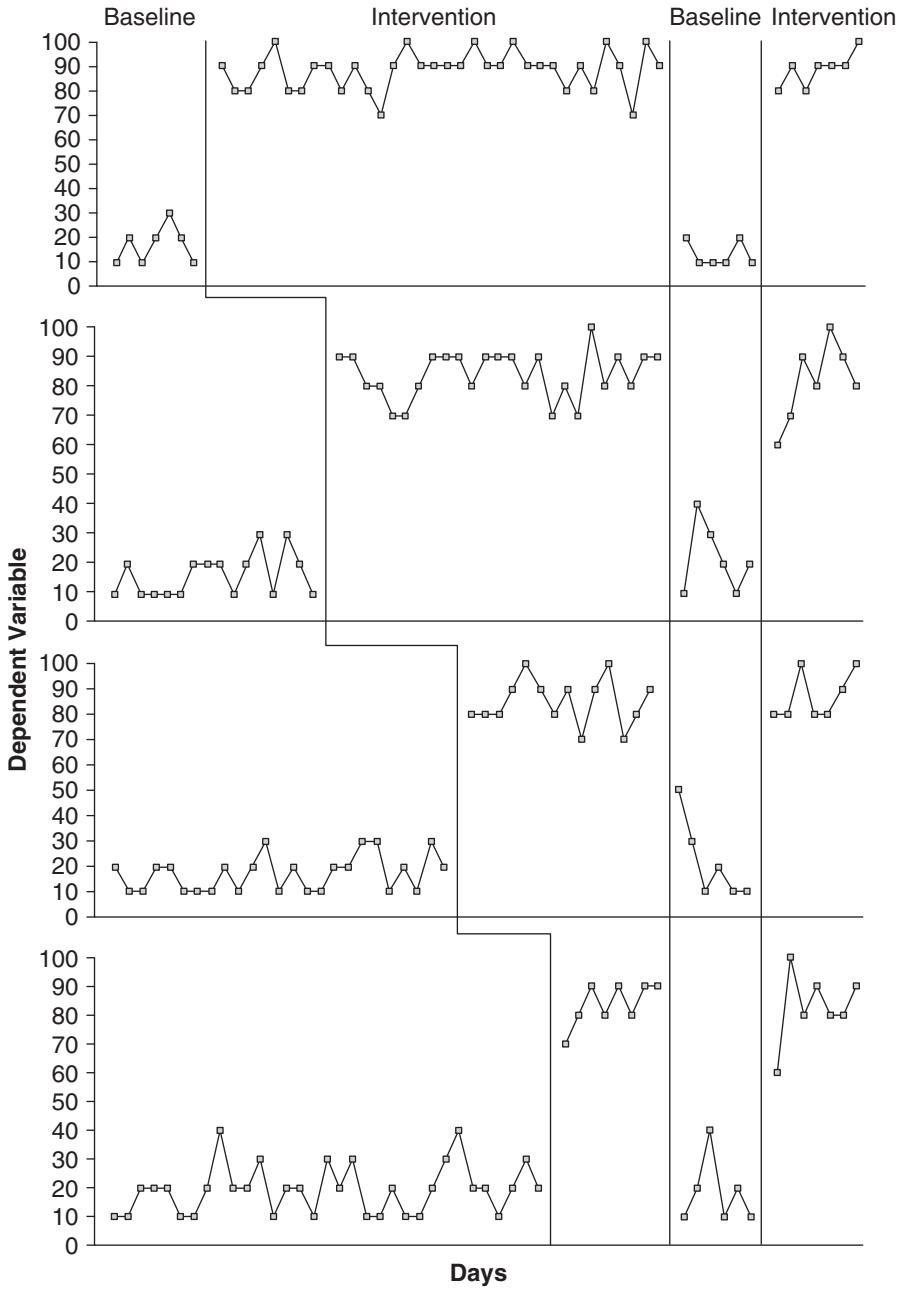



**Figure 5** Simulated Combination of Multiple Baselines Across Students Design and Reverse Design

the actual pattern observed after manipulation of the independent variable.

In addition to comparing the level, trend and variability, the researcher also examines the *immediacy of effect, overlap,* and *consistency of data in similar phases*. *Immediacy of effect* refers to the level of change between the last three data points in one phase and the first three data points of the next. The more rapid, or immediate, the effect, the more likely the effect is caused by manipulations of the independent variable. *Overlap* refers to the proportion of data from the second phase that overlaps with data from the previous phase along the ordinate, or vertical, axis. The smaller the proportion of

overlapping data points (or conversely, the larger the separation), the more likely the data document a change caused by the intervention. *Consistency of data in similar phases* involves looking at data from all baseline phases and from all intervention phases as well as examining the extent to which there are consistent patterns. The greater the consistency, the more likely the data represent a functional effect caused by the intervention.

Regardless of the single-case research design used in a study, a visual analysis of (a) level, (b) trend, (c) variability, (d) overlap, (e) immediacy of effect, and (f) consistency of data patterns across similar phases is used to assess whether the data demonstrate at least three indications of an effect across at least three points in time. If this criterion is met, then the data are deemed to document a functional relation, and an inference might be made that change in the dependent variable is causally related to the manipulation of the independent variable.

## External Validity

Although demonstration of experimental effects is possible, such results might have limited relevance for individuals and contexts outside those used in a given study. *External validity* refers to the extent to which findings from a study have relevance beyond the specific parameters of the study (e.g., "Do the findings have value for considering intervention with other participants, in other locations, by other interventionists?").

The external validity of single-case research lies in the replication of findings across multiple studies. If five or more studies produce similar findings and are conducted with different individuals across a range of contexts, then confidence in the generality of the findings is increased. Just as the replication of the effect *within* the design is an essential component of single-case research studies, the replication of findings *across* the design is another core feature of this approach to science.

## Use of Single-Case Research to Define Evidence-Based Practices

A major development in education and psychology is the identification of evidence-based practices. Both federal and state agencies are now expected to identify practices that are not just "possible" but empirically documented to benefit consumers (students, families, and patients). This expectation establishes a need for research standards that will be used to document a practice as "promising" or "evidence based." Single-case research holds promise as one approach for documenting evidence-based practices, and standards for using single-case research for this purpose have been proposed by the American Psychological Association and the Council for Exceptional Children. In each case, the expectation is established that a group of studies employing the same practice (intervention) must document experimental control, be conducted across multiple researchers, include 20 or more combined participants, and not include demonstrations of treatment failure. It can be expected that a major contribution of single-case research designs will be the future identification of evidence-based practices.

*Robert H. Horner and Scott A. Spaulding*

***See also*** Dependent Variable; Effect Size, Measures of; Independent Variable; Variability, Measure of

### Further Readings

American Psychological Association. (2002). Criteria for evaluating treatment guidelines. *American Psychologist, 57,* 1052–1059.

Hersen, M., & Barlow, D. H. (1976). *Single-case experimental designs: Strategies for studying behavior change.* New York: Pergamon.

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children, 71,* 165–179.

Kennedy, C. (2005). *Single-case designs for educational research.* Boston: Allyn and Bacon.

Parker, R., & Hagan-Burke, S. (2007). Useful effect size interpretations for single case research. *Behavior Therapy, 38,* 95–105.

Parsonson, B., & Baer, D. (1978). The analysis and presentation of graphic data. In T. Kratochwill (Ed.), *Single-subject research: Strategies for evaluating change* (pp. 105–165). New York: Academic Press.

Tawney, J. W., & Gast, D. L. (1984). *Single-subject research in special education.* Columbus, OH: Merrill.

Wolery, M., & Dunlap. G. (2001). Reporting on studies using single-subject experimental methods. *Journal of Early Intervention, 24,* 85–89.

# Social Desirability

Social desirability is the tendency for research participants to attempt to act in ways that make them seem desirable to other people. Such attempts to "look good to others" can compromise the validity of research, particularly research with participants who know they are being studied. Frequently viewed as a response bias, social desirability is often associated with self-report questionnaires; however, it can also affect any research based on behavioral observation. Researchers from many sciences, including psychology, business, public opinion, medicine, political science, sociology, and exercise science, must consider the potential effects of social desirability. Researchers have long studied the nature of social desirability, the factors affecting social desirability, its potential impact on research, and various methods for coping with the problem.

## The Nature of Social Desirability

Seeming simple at first glance, social desirability has a rather complex nature. There are at least two important issues concerning the nature of social desirability. The first issue is its psychological content—the number and nature of the psychological dimensions likely to be affected by social desirability. Early research focused on a single psychological dimension—the need for approval from others—as the basis of social desirability. That is, early work focused on the general tendency to claim or demonstrate desirable characteristics and deny or conceal undesirable characteristics (i.e., to claim an uncommon degree of virtue and deny a common degree of vice). More recent work, most notably research conducted by Del Paulhus and his colleagues, has highlighted two dimensions of social desirability—an *egoistic* dimension reflecting the exaggeration of traits such as competence, dominance, and intelligence, and a *moralistic* dimension reflecting the exaggeration of traits such as friendliness, impulse control, and responsibility.

A second issue concerning the nature of social desirability is intentionality. Although socially desirable behavior is sometimes conscious and intentional, it also can be unconscious or unintentional. *Impression management* occurs when people intentionally exaggerate socially desirable qualities (or intentionally deny socially undesirable qualities). *Self-deception* occurs when people unintentionally exaggerate such qualities. That is, self-deception occurs when people truly but inaccurately believe that they have more desirable qualities than they actually have. This issue is closely related to a common terminological distinction between response sets and response styles. Although researchers are inconsistent in using these terms, a *response set* is a temporary tendency toward social desirability (e.g., impression management that occurs while completing a job application), and a *response style* is an enduring tendency for some people to respond in a socially desirable manner (e.g., a stable disposition for some people to view themselves in an overly positive light).

## What Produces Social Desirability

Several factors affect socially desirable behavior. First, some people might be predisposed toward socially desirable behavior. Such people are said to have a socially desirable response style, behaving in a socially desirable way in many kinds of research settings.

A second factor affecting social desirability is the behavioral characteristic being examined in a given study. That is, some behavioral characteristics are more likely to elicit socially desirable behavior than are others. For example, some traits, behaviors, or attitudes are particularly valued, stigmatized, or otherwise sensitive (e.g., mental health, racial attitudes, sexual practices, drug use, morality, and general psychological adjustment). When such characteristics are measured and/or observed, participants' responses are relatively likely to be affected by social desirability. For example, whether they are intentionally attempting to present an image consistent with common social norms or they are unintentionally failing to recognize their true nature, some people might claim racial tolerance when they are, in fact, racially intolerant.

A third factor affecting socially desirable responding is the observational context; socially desirable responding is more likely to occur in

some measurement contexts than in others. Most simply, desirable responding is more likely to occur when people expect consequences from their responses than when they expect no consequences. For example, a person completing an integrity survey as part of a job application might be more likely to exaggerate his or her virtues than a person completing the same survey as an anonymous participant in a large research project. Interestingly, some observational contexts elicit socially *un*desirable responding rather than socially desirable responding—a phenomenon called *malingering*. Specifically, when people can benefit from being viewed as suffering from psychological, social, or physical deficiencies (e.g., when suing for cognitive impairment suffered during an accident), they might respond in a way that exaggerates these deficiencies or problems.

## Impact of Social Desirability

Social desirability is often recognized as a bias that creates problems for research and for applied measurement. Most directly, social desirability can compromise the validity of scores on a measure. That is, if peoples' measured behaviors or responses are affected by social desirability, then those measurements are biased as indicators of their intended construct. For example, a researcher wishes to measure participants' self-esteem by using a self-report questionnaire. Ideally, people who obtain high scores on the questionnaire would have high levels of self-esteem. However, if some participants have a socially desirable response style (e.g., an egoistic bias), then those people might respond to the questionnaire in a way that exaggerates their true level of self-esteem. Consequently, at least some high scores on the questionnaire reflect social desirability motivation rather than self-esteem; thus, the validity of the responses as indicators of self-esteem has been compromised. When a measure is associated significantly with social desirability, researchers sometimes say that the instrument is "contaminated" by social desirability.

Compromised measurement validity can have harmful effects on research. For example, social desirability bias might produce spuriously strong correlations, incorrectly indicating that two variables are more strongly associated with each other than they truly are. Again, imagine a researcher interested in self-esteem and its association with creativity. Furthermore, imagine that, in reality, the association between the two constructs is weak. Unfortunately, socially desirable response styles might produce a spuriously strong correlation between measures of the constructs. This arises from the combined effect of two factors. First, participants might differ in the tendency to respond in a socially desirable manner (i.e., some people have a socially desirable response style and some do not). Second, the two constructs might be affected by social desirability. That is, both self-esteem and creativity might be desirable qualities, eliciting overly positive responses from some participants but not others. The fact that people with a socially desirable response style would obtain inflated scores on *both* measures will produce a relatively strong correlation between the measures, not reflecting the truly weak association between self-esteem and creativity. Although the real correlation between the traits of self-esteem and creativity might be weak, social desirability bias can affect scores on measures of both constructs and inflate the correlation between the measures. Consequently, the researcher is likely to conclude mistakenly that self-esteem is strongly related to creativity.

Measurement validity that is compromised by social desirability can also have harmful effects in applied measurement. When decisions are even partially based on observations of behavior (including verbal behavior and self-report behavior), the quality of those decisions hinges on the validity of the observations. As a factor that might diminish measurement validity, social desirability can reduce the quality or accuracy of decisions. For example, hiring is a real-life decision that might be based partly on some form of behavioral observation (e.g., interview and personality tests). Consequently, if applicants engage in substantially biased impression management, then hiring decisions might result in the employment of relatively less-qualified applicants.

## Coping With Social Desirability

In attempting to cope with potential problems introduced by social desirability, researchers have

developed strategies targeting two main goals. First are strategies designed to prevent or minimize the occurrence of socially desirable responding. Among these are strategies for managing the content of psychological surveys and questionnaires. For example, researchers can use relatively neutral or "unloaded" phrasings for questionnaire items—an item like "I am very irresponsible" might be rephrased to "Sometimes, I am not as responsible as other people," which is a less undesirable (i.e., less evaluatively loaded) phrasing that people might be more willing to endorse. In managing test content, researchers can also use a forced-choice format, in which participants are presented with two equally desirable (or undesirable) characteristics (e.g., "responsible" and "creative") and must choose the one that is more descriptive of themselves. Although some research suggests that this is not a fully effective method for controlling impression management, a forced-choice format is often believed to minimize social desirability. Also among the strategies for preventing or minimizing the existence of bias are methods of managing the testing context. As mentioned earlier, facets of the testing context can affect the occurrence of socially desirable behavior; thus, researchers can emphasize participants' anonymity when possible, or they can warn participants that inaccurate or deceptive behavior can be detected in some way (whether such detections can be made or not).

A second set of strategies has been developed to detect the existence of biased responding and intervene in some way. Several such approaches have been developed, but the most common is the use of self-report scales intended to detect socially desirable responding. For example, some wide-ranging, multidimensional inventories such as the well-known Minnesota Multiphasic Personality Inventory include "validity" scales intended to detect various forms of response bias. Similarly, specialized questionnaires, such as the Marlowe-Crowne Social Desirability Scale and the Brief Inventory of Desirable Responding (BIDR), have been developed as stand-alone inventories that can be used as part of larger projects in which researchers are concerned about social desirability. Such scales often include items reflecting qualities so rare (e.g., "I have never told a lie") that few people could legitimately endorse them. Participants who endorse many such items are assumed to be presenting an inaccurately positive image of themselves.

After socially desirable behavior has been detected, researchers can intervene in some way. For example, researchers might omit data from participants who have provided questionable responses, as indicated by their scores on a measure of socially desirable responding. That is, analyses can proceed without data that are potentially biased. Alternatively, researchers might statistically control for scores on a social desirability scale. Statistical procedures such as partial correlation or multiple regression could be used in this way. For example, a researcher might compute the partial correlation between self-esteem scores and creativity scores, controlling for scores on the BIDR. Such procedures statistically account for the fact that some people might have responded with a socially desirable response style.

## Bias Versus Substance

A final important issue is whether social desirability reflects inaccuracy/bias or valid responding. The typical perspective is that social desirability is a source of error, bias, or contamination compromising the quality of social and behavioral research. However, some researchers maintain a different perspective, offering various arguments that social desirability has some legitimately substantive psychological meaning. For example, some researchers have found that statistically correcting for the potential bias caused by social desirability does *not* affect the association between personality trait scores and criteria such as trait ratings from well-acquainted informants. Similarly, some research suggests that people who obtain high scores on measures of social desirability are actually viewed by others in desirable ways (e.g., they are recognized as relatively well-adjusted). Such findings suggest that high scores on measures of social desirability reflect, to some degree, validly desirable personality characteristics.

Despite such emerging evidence, most researchers still view social desirability as a source of error, bias, and invalidity in research. After decades of concern over social desirability and after decades of empirical research regarding the existence and potential effects of social desirability, our

understanding of the nature, impact, and control of social desirability continues to evolve.

*R. Michael Furr*

***See also*** Confounding; Construct Validity; Content Validity; Criterion Validity; Internal Validity; Interviewing; Partial Correlation; Psychometrics; Reliability; Response Bias; Statistical Control; Survey; Validity of Measurement

**Further Readings**

Block, J. (1965). *The challenge of response sets*. New York: Appleton-Century-Crofts.

Crowne, D. P., & Marlowe, D. (1964). *The approval motive*. New York: Wiley.

Edwards, A. L. (1957). *The social desirability variable in personality assessment and research*. New York: Dryden.

Furr, R. M., & Bacharach, V. R. (2008). *Psychometrics: An introduction*. Thousand Oaks, CA: Sage.

McCrae, R. R., & Costa, P. T., Jr. (1983). Social desirability scales: More substance than style. *Journal of Consulting and Clinical Psychology, 52,* 882–888.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp.17–59). New York: Academic Press.

Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. In H. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 67–88). Hillsdale, NJ: Lawrence Erlbaum.

# Software, Free

The Internet offers a large and growing number of free computer programs of good quality that can be used in the planning, performance, and analysis of research studies. These applications can meet many, most, or all the needs of most researchers. Some programs can be downloaded and run on one's own computer, and some are online programs (i.e., interactive statistical calculating pages on the Internet). Some of these interactive pages are also downloadable. Most of the downloadable programs are suitable for personal computers (PCs), and some are made for Apple-based Macintoshes (Macs). Some are Windows programs, some of the older ones are DOS programs, and some will run in Linux and other operating systems. Some can be run in hand-held computers.

Some free programs or program packages provide a wide variety of procedures, whereas others are very specialized (i.e., they perform limited tasks or are geared toward the needs of narrow research fields).

Selected programs are mentioned next. The websites where they are available can be identified via a computer search engine or published annotated lists of free programs.

## Multipurpose Software

The following selected multipurpose programs or packages are listed alphabetically.

*EasyReg* is an extremely versatile program designed for econometric studies. It can read data files and spreadsheets, manipulate and summarize data, and perform numerous complex analyses with simple point-and-click operation. Guided tours that explain the methods are provided.

*Episheet* is a spreadsheet that performs many procedures commonly used in epidemiological studies.

*MacAnova,* which has Mac, Windows, DOS, Linux, and UNIX versions, performs numerous simple and complex analyses. It can read data files and manipulate data. Data can be pasted. *MacAnova* is primarily command driven (i.e., instructions must be typed in). Full documentation is available.

*OpenEp*i is an interactive page providing programs for use in public health and medicine. It is compatible with PC, Mac, and Linux browsers, and it is downloadable. It provides statistics for counts and person-time rates in descriptive and analytic studies; stratified analysis with exact confidence limits; matched pair analysis; sample size and power calculations; random numbers; chi-square for dose-response trend, sensitivity, specificity, and other evaluation statistics; and $R \times C$ tables. *OpenEpi* requires the keyboard entry of data that have already been counted or summarized. The program is available in four languages.

*OpenStat* was written as an aid to students and teachers for introductory and advanced statistics courses, primarily in the social sciences, and it is

said to perform nearly all the analyses required in statistics courses. It is available in Windows and Linux versions. A manual is provided. Raw data can be entered to create a data file, or it can use data files created by other programs (e.g., *EpiData*).

The *R Project for Statistical Computing* provides many procedures based on R (a language and environment for statistical computing and graphics: data handling and storage, tools for data analysis, and graphic facilities). The procedures can be incorporated in a broader data-analysis and presentation tool, which is tailored to meet specific needs. There are Mac, Windows, DOS, Linux, and UNIX versions. Programs are versatile but require typing of instructions; the language has to be learned. Manuals and guides are available.

*Simple Interactive Statistical Analysis* (*SISA*) performs simple calculations online. Most of the modules are downloadable. *SISA* requires the entry of data that have already been counted or summarized.

The *VassarStat* website provides several small programs for the online analysis of proportions, ordinal, categorical, and numerical data; analysis of variance and covariance; and other procedures. Data might be entered at the keyboard or pasted.

*Wessa.net* performs calculations online. It has modules that provide descriptive statistics, multiple regression, hypothesis testing, plotting of mathematical equations, and other procedures. Data might be entered at the keyboard or pasted.

*WinIDAMS* was developed by UNESCO to provide a reasonably comprehensive data management and statistical analysis software package. Its functions include regression analysis, analysis of variance, discriminant analysis, cluster analysis, and principal component analysis. Guides and a manual are available. Raw data can be entered into WinIDAMS to create a data set, or it can use data files created by other programs .

*WinPepi,* or *Pepi*-for-Windows, is a suite of Windows-based programs for epidemiologists, containing more than 100 modules, offering most of the statistical procedures commonly used in epidemiological studies (with some notable exceptions, such as Cox regression analysis), as well as many that are less commonly used or not easily found (or not easily found in a user-friendly format). The programs are easy to use and have comprehensive manuals. A desktop portal provides easy access to an alphabetical list of procedures and to the programs and manuals. In general, *WinPepi* requires the entry of data that have already been counted or summarized, either manually or by using a program (e.g., *EpiData*) that processes primary data. Data are entered at the keyboard or pasted from a file or spreadsheet .

Three of the programs—*Episheet, OpenEpi*, and *WinPepi*—can be regarded as specifically epidemiological software, in that they use epidemiological language and give emphasis to procedures and outputs that are relevant to studies of health and disease. But they can be used in other research contexts also, provided that allowance is made for their epidemiological labels (e.g., their naming the samples to be compared "cases" and "controls," or "exposed" and "nonexposed").

## Planning and Preparatory Phases of a Study

Estimates of required sample sizes and power, for studies of various kinds, can be provided by *WinPepi* and several other programs, such as *PS, Epi Winscope, Csurvey,* and (online) *OpenEpi* and *Lenth's Java applets for power and sample size*. The number and size of clusters required for cluster-randomized trials are computed by *WinPepi* and *Sampsize*.

Programs that select random samples include *WinPepi, Survey Toolbox* (which is geared to veterinary epidemiological studies), and (online) by *OpenEpi, SISA, Research Randomizer*, and the *VassarStat* page.

Randomization (simple, blocked, and balanced assignment of subjects to the arms of trials), random sequencing, and an aid to minimization are provided by *WinPepi*. *Randomization.com* and *SISA* can perform randomization online.

### *Data Collection*

*EpiData* and *Epi Info* can streamline the data-collection procedure. They produce data entry forms, such as questionnaires, from sets of typed questions or other text, using conventions for numeric, yes/no, and other data fields. The programs allocate variable names, field lengths, and so on, using information derived from the form. The form is subsequently shown on a computer

screen so that, for each subject in turn, data can be entered and stored in a data file created by the program; this database can be exported to other programs. The forms can be printed for use as paper questionnaires or displayed in computer-assisted interviews. Data-entry constraints (checks for acceptable ranges of responses, logical consistency between responses, etc.) and skip patterns (e.g., jumping over "female" questions if the subject is a male) can be built in, and new variables can be constructed (calculation of age, summation of scales, etc.). To ensure a high quality of data, the data can be entered twice, and the program will compare the two sets of data. The programs can also count the data and prepare tabulations (which can be used in statistical programs that require entry of counts or tabulated data), and can perform calculations. Detailed guides are available. *EpiData* is available in more than ten languages, and *Epi Info* is available in four.

*EpiSurveyor* performs similar functions for hand-held devices, creating data entry forms and enabling their use in the field and the subsequent transfer of data to a desktop or laptop computer for analysis.

*IMPS* (the Integrated Microcomputer Processing System), developed by the U.S. Census Bureau, is said to allow persons with little or no computer experience to contribute to the data processing operation. It deals with (inter alia) data entry, checking, editing, and tabulation.

Programs to facilitate the performance of Web-based research, including the creation of data entry forms, include *Web Survey Toolbox* and *LimeSurvey. QPL* (Questionnaire Programming Language) is a Web questionnaire development and content analysis system—using simple language to describe the questionnaire, it is said to automatically build all the files needed to create a dynamic website.

Many of the statistical programs will accept data pasted from text files or spreadsheets. A large number of free text editors and spreadsheets (such as the *Sphygmic Software* spreadsheet) are available. *Editpad Lite, Crimson Editor,* and *Notetab Light* are particularly useful text editors for this purpose, because they permit rectangular selections—that is, rectangular blocks of data (such as selected groups of adjacent columns) can be copied for pasting into the statistical program.

### *Data Management*

As noted previously, *EpiData, Epi Info,* and *WinIDAMS* can create data sets. *EpiData* and the *Epi Info* software permit checks on consistency, corrections, recoding, the creation of new variables, the splitting and merging of data sets, and conversion of the data to formats appropriate for various statistical programs; the database can be encrypted. *WinIDAMS* offers facilities for sorting and merging files, data editing, checking of codes and consistencies, correcting, listing, subsetting, aggregating, and merging and transforming data, including the construction of new variables.

*Amelia* and *NORM* can be used to replace missing values with imputed ones.

## Analysis

The wide range of statistical analyses that can be performed by free multipurpose packages and other free programs, and the wide range of the software that can perform these analyses, is illustrated by the following list of selected procedures. Emphasis is here given to uses in epidemiological research.

Disease clustering can be explored by *CLUSTER,* a Centers for Disease Control and Prevention program that uses 12 statistical methods to evaluate time, space, and time-and-space clustering.

Seasonal variation can be appraised by *Episheet* and *WinPepi.*

Analyses of stratified data in two samples (independent or matched) can be performed by *WinPepi,* for categorical (dichotomous, nominal, or ordinal) and numerical data. *Episheet, Epi Info, VassarStat,* and *Win Episcope* can perform such analyses for dichotomous data.

Multiple linear regression can be performed by *WinIDAMS, WinPepi, OpenStat, MacAnova,* and *EasyReg,* as well as (online) by *Wessa.net.*

Multiple logistic regression can be performed in Windows by *WinPepi, Epi Info, MacAnova,* and *OpenStat.* In DOS, there are *MultLR, Logistic* (unconditional logistic regression), and *Clogistic* (conditional logistic regression). John C. Pezzulo and Kevin Sullivan's "*Logistic Regression*" page conducts multiple logistic regression online.

Cox regression analysis can be performed by *OpenStat, Epi Info, MacAnova,* and *CoxSurv* (a

DOS program). Pezzullo and Sullivan's *Cox Proportional Hazards Survival Regression* page does it online.

Multiple Poisson regression can be performed by *EasyReg, MacAnova,* and *WinPepi.*

*WinPepi* can provide a comprehensive assessment of the validity of screening or diagnostic tests or other measures that have yes/no results, and also of tests that have a range of values; it produces ROC curves and can compare tests. ROC curves are examined by *VassarStat* and, in detail, by *ROC Analysis* and *GraphROC. OpenEpi* assesses yes/no tests and tests with a range of values, and produces a ROC curve. Yes/no tests are appraised by *Win Episcope* and (online) by *On-Line Clinical Calculator.*

The reliability of measurements can be appraised by *WinPepi,* which computes *kappa* and related measures of the reliability of categorical data, as well as numerous coefficients that express the reliability of numerical measurements. *DagStat* computes *kappa* and related measures of the reliability of categorical data. Online, *kappa* for two or more categories is computed by *VassarStat* and *QuickCalcs.*

An assessment of scales (their internal consistency—that is, Guttman scalability and Cronbach's alpha—and their discriminatory capacity) can be performed by *WinPepi* and *OpenStat.*

Studies of child growth can use *Epi Info* or WHO's *Anthro* program to make comparisons with standards and compute percentiles or $z$ scores.

A capture-recapture procedure is used by *WinPepi* to estimate the number of cases of a disease in a population from incomplete overlapping lists.

Survival analysis is performed by *WinPepi, Epi Info, OpenStat, Easyreg, VassarStat, Episheet,* and *LaMorte,* and (in DOS) by *KMsurv. Hutchon's interactive page* does survival analysis online.

*WinPepi* performs direct standardization of rates, proportions, means, and other statistics, including the standardization method that gives each year of age the same weight, and the computation of standardized morbidity or mortality ratios. *Episheet* and *LaMorte* can do standardization of rates and proportions. *PamComp, OpenEpi,* and *LaMorte* calculate standardized morbidity or mortality ratios.

Cluster samples can be analyzed by *WinPepi, Epi Info,* and *WinIDAMS.*

Multilevel studies can be analyzed by *MLwiN* and by *MIXOR, MIXREG, MIXNO,* and *MIXPREG,* which perform multilevel analysis for linear, logistic, probit, and Poisson regression and for survival analysis.

## Meta-Analysis

*WinPepi* can compare the results of different studies, providing tests and quantitative measures of heterogeneity, and can combine the results, using fixed-effect and random-effects models. The analysis might be based on odds ratios, rate or risk ratios, differences between means, standardized mean differences (effect sizes), ratios of means, simple proportions, $p$ values, *kappa,* indices of the validity of screening or diagnostic tests, and other measures. It calculates the fail-safe $N$, and provides tests for a skewed funnel plot that suggests publication bias, and a sensitivity analysis showing the effect of excluding each study in turn.

*MIX* and *Episheet* use spreadsheets to perform meta-analyses.

DOS-based packages that perform meta-analyses include *Meta-analysis, EpiMeta,* and *EasyMA.*

## Evaluation of the Results

The effect of misclassification on an observed rate or proportion or odds or risk ratio can be appraised by *WinPepi*.

The effect of regression to the mean, in longitudinal studies in which subjects were selected because of their extreme values, can be appraised or neutralized by *WinPepi,* which can also use analysis of covariance to avoid it, as can *OpenStat* and *VassarStat.*

The adjustment of $p$ values to take account of the performance of multiple tests on the same data is performed by *Multi* and *WinPepi.*

Statistical synergism (additive or multiplicative) and antagonism, and evidence for a cohort effect (using median or mean polish) are appraised by *WinPepi. Epinetcalculation.xls* also computes measures of synergism, using a spreadsheet.

A sensitivity analysis can be conducted by *WinPepi* to examine the effect of hypothetical unmeasured confounders on an odds or hazard ratio.

The clinical or public health significance of study findings is examined by most of the relevant

epidemiologic programs: Most programs that analyze clinical or prophylactic trials or analytic epidemiological studies compute the number needed to treat to avoid one case (or similar measures), and attributable, prevented, and preventable fractions in the population; and most of the programs that appraise screening and diagnostic tests and risk markers examine their clinical value as reflected in, for example, predictive values.

### Preparation of the Report

*Epi Info* produces simple tables, including cross-tabulations, in a format suitable for publishing. It can also create maps.

Numerous free programs can draw graphs. The following are examples:

*Epi Info* and *EpiData* draw graphs.

*RJSgraph* and *SpectrumViewer* draw excellent data-plot graphs (Y-plots and XY-plots), with many optional trimmings; they can transform data, provide statistics, and so on. *RJSgraph* can plot equations and provides a range of regression lines.

The *Sphygmic Software* and *Open Office* spreadsheets draw pie, bar, line, scatter, 3-D, and other charts.

*Dataplot* is a multi-platform (Windows, UNIX, VMS, Linux, etc.) software system for scientific visualization, statistical analysis, and nonlinear modeling. It supplies XY plots, distributional, time series, 3-D, multivariate, quality control, and numerous other plots.

*MacAnova* can draw scatter plots, line plots, box plots, histograms, stem- and-leaf displays, and other graphs.

*OpenStat* can draw box plots, stem-and-leaf plots, and pie and other charts.

*Kyplot* (version 2) is an elaborate program that draws line, scatter, bar, pie, 3-D, and many other charts, as well as performing statistical analyses.

*Wessa.net* draws histograms and box, stem-and-leaf, and other plots.

*SBHisto* and *VassarStat* draw simple histograms.

*SSP* (Smith's Statistical Package) draws simple histograms, bar charts, time-series graphs, scatterplots, and box-and-whiskers diagrams.

*WinPepi* draws time-series and box-and-whisker diagrams, ROC curves, epidemic curves, and other diagrams.

*Diagrams* and *EpiGram* draw flowcharts and other diagrams.

*Flowchart.com* draws flowcharts online.

*WinIDAMS* draws scatterplots.

*Episheet* draws survival curves and forest plots.

### Conclusion

Free software of good quality is of special value to students and to investigators in less well-developed countries and other impecunious researchers. But apart from the question of cost, the programs tend to be relatively easy to use. This can in many cases make free software more attractive than widely used and elaborate commercial statistical packages that demand a great deal of know-how—often necessitating considerable study, or participation in special courses—and experience. Even competent users of elaborate commercial packages might find that for some purposes, free programs are useful substitutes or supplements.

*J. H. Abramson*

***See also*** Meta-Analysis; R; Regression to the Mean; WinPepi

#### Further Readings

Abramson, J. H., & Abramson, Z. H. (2008). Appendix C. Free computer programs. In *Research methods in community medicine: surveys, epidemiological research, programme evaluation, clinical trials* (6th ed.) (pp. 389–405). Chichester, UK: Wiley.

Pezzullo, J. C. (2008). *Free statistical software*. Retrieved October 15, 2008, from http://StatPages.org/javasta2.html

Pezzullo, J. C. (2008). *Web pages that perform statistical calculations*. Retrieved October 15, 2008, from http://statpages.org

## Spearman–Brown Prophecy Formula

Charles Spearman and William Brown separately derived a formula to predict the reliability of a test

when its length was altered by the addition or subtraction of parallel items. Each presented their formula in 1910 in Volume Three of *The British Journal of Psychology*. Because their articles were published at the same time, the formula is known as the Spearman–Brown prophecy formula.

The basic premise of classical test theory is that an observed test score consists of the sum of the following two components: true score and error score. For any individual examinee, the two cannot be separated, but for a group of examinees, the variance attributed to each source can be estimated. Test reliability, which is an important characteristic of test score quality, is defined as the ratio of true score variance to observed score variance.

By definition, error scores are random perturbations—they covary zero with each other as well as with true scores. It is well established within statistics that the variance of a sum is equal to the sum of the variances plus two times the covariances.

Consider a test to which items are added that measure the same construct equally as well as the items already in the test (parallel items). When an item is added, the true score variance of the test is increased by the true score variance of the item as well as two times the sum of true score covariances. The error variance of the test is increased by the item error score variance, but because error scores have zero covariance with everything else (by definition), the true score variance of a test increases more rapidly than the error score variance. Thus, adding parallel items increases the reliability of a test.

An early important use of the prophecy formula allowed the estimation of score reliability from a single test administration. For example, when using the split-half reliability calculation method, a single test is divided into two halves with equivalent properties and the reliability of the half tests is calculated by correlating the two. The Spearman–Brown formula allows the test developer to estimate the reliability of the full-length test from the reliability of the test halves. Additionally, the formula allows estimation of full-length test reliability when any number of individual items are added to or subtracted from a test.

The Spearman–Brown formula relies on the assumption of strictly parallel tests or test items. Parallel tests or items are those that have identical properties, measure the same constructs, and have the same level of difficulty. The Spearman–Brown prophecy formula will only produce a valid estimation if the change in the length of the test is derived from parallel components. If an increase in test length is accomplished by adding items of a lesser quality, the Spearman–Brown formula will overestimate the reliability of the longer test. If the new items are statistically better than the average items of the original test, then Spearman–Brown will underestimate the reliability.

The Spearman–Brown prophecy is expressed in the following equation:

$$\hat{r}_{tt} = \frac{kr_{AB}}{1 + (k-1)r_{AB}},$$

where $k$ is the factor by which the length of the test is changed or the ratio of the new test length to the old test length, $r_{AB}$ is the reliability of the original test, and $\hat{r}_{tt}$ is the estimated reliability of the test $k$ times as long as the original test.

When a test developer wishes to estimate reliability from a test that is doubled in length or when $k = 2$, the equation simplifies to

$$\hat{r}_{tt} = \frac{2r_{AB}}{1 + r_{AB}}.$$

Consider a researcher who wishes to estimate the impact of adding 30 additional equivalent items to a 20-item test that has yielded a reliability of .50. The new 50-item test will be 2.5 times as long as the old test; it will increase by a factor of 2.5. Using the Spearman–Brown prophecy formula, the answer is

$$\hat{r}_{tt} = \frac{2.5(.50)}{1 + 1.5(.50)} = \frac{1.25}{1.75} = .71.$$

The estimated reliability of the new test is .71.

Another form of the equation allows the estimation of the number of items needed to achieve a desired level of reliability of test scores when the original test reliability is known. When solved for $k$, the Spearman–Brown formula is

$$k = \frac{\hat{r}_{tt}(1 - r_{AB})}{r_{AB}(1 - \hat{r}_{tt})}.$$

The resulting $k$ equals the factor by which the test should be changed to achieve a desired full test reliability of $\hat{r}_{tt}$.

Suppose that a 20-item math test is producing a reliability of .60; how many comparable items would need to be added to the test to achieve a reliability of at least .80? Using the form of the equation that solves for the factor $k$, the answer is

$$k = \frac{\hat{r}_{tt}(1 - r_{AB})}{r_{AB}(1 - \hat{r}_{tt})} = \frac{.80(1 - .60)}{.60(1 - .80)} = \frac{.32}{.12} = 2.67.$$

The test would need to be increased by a factor of 2.67. Taking the factor of 2.67 times the original test length of 20 items requires a total of 54 items, so 34 items would need to be added to achieve a reliability of .80.

Practical considerations must be noted when using the Spearman–Brown formula. The formula demonstrates that net change in reliability is a function of the reliability of the original test. Adding items to a test that begins with low reliability will yield a larger change in reliability than a test that begins with a higher reliability. Thus, as the test reaches higher levels of reliability, the addition of parallel items yields diminishing returns.

*Neal Kingston and Gail Tiemann*

***See also*** Classical Test Theory; Error; Reliability; Split-Half Reliability; True Score

**Further Readings**

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3,* 296–322.

Spearman, C. (1910). Correlation calculated with faulty data. *British Journal of Psychology, 3,* 271–295.

# Spearman Rank Order Correlation

For ordinal-level data, the Spearman rank order correlation is one of the most common methods to measure the direction and strength of the association between two variables. First put forth by British psychologist Charles E. Spearman in a 1904 paper, the nonparametric (i.e., not based on a standard distribution) statistic is computed from the sequential arrangement of the data rather than the actual data values themselves. The Spearman rank order correlation is a specialized case of the Pearson product-moment correlation that is adjusted for data in ranked form (i.e., ordinal level) rather than interval or ratio scale. It is most suitable for data that do not meet the criteria for the Pearson product-moment correlation coefficient (or Pearson's $r$), such as variables with a non-normal distribution (e.g., highly skewed) or that demonstrate a somewhat nonlinear tendency.

Although Pearson's $r$ indicates the strength of the linear relationship between two variables, the Spearman rank order correlation shows the strength of the monotone associations. A monotonic relationship is one in which all $x$ and $y$ variables are arranged in ascending order and are compared for their differences in ranks. In other words, the statistic determines to what degree one data set influences another data set. An increasingly positive monotonic relationship would be one in which as $x$ increases in rank (and value), $y$ also increases (or stays the same) in rank, and the data pairs are concordant with one another. In contrast, an increasingly negative monotonic relationship exists when as $x$ increases in rank (and value), $y$ decreases (or stays the same) in rank. In this case, the data pairs are discordant with one another and exhibit large differences between their respective ranks. In contrast to Pearson's $r$, the Spearman rank order correlation does not differentiate between linear and monotonic associations because of the ordinal level scale of the data. This property enables the Spearman rank order correlation to be a suitable nonparametric alternative to the Pearson's $r$ when assumptions regarding linearity cannot be met or are unknown.

This entry discusses several aspects of the Spearman rank order correlation, including methods for computing, the influence of tied rankings, and statistical significance and significance testing.

## Computation

The Spearman rank order correlation measures the degree of association of ordinal-level data by examining the ratio of the sum of the squared differences in the ranks of the paired data values to the number of variable pairs. Computationally, the Spearman rank correlation coefficient ($r_s$) is defined by the formula

$$r_s = 1 - \frac{6\sum d^2}{N(N^2 - 1)},$$

where $d$ is the difference in statistical ranks between the paired variables $x$ and $y$, $\Sigma d^2$ is the sum of the squared differences between the ranks of the paired variables, and $N$ is the number of paired data values. As with the Pearson's $r$, values of $r_s$ range from a minimum of $-1.0$ to a maximum of 1.0. Increasingly negative $r_s$ values indicates a negative monotonic relationship between the ordered pairs, and $x$ and $y$ are inversely related to one another (i.e., as one variable increases, the other tends to decrease). Conversely, increasingly positive $r_s$ values indicates a positive monotonic relationship, and $x$ and $y$ covary in the same direction (i.e., as one variable decreases, the other is apt to also decrease). The closer the Spearman rank correlation coefficient is to the extremes (1.0 or $-1.0$), the stronger the association between the variables. If no association exists between the variables, then $r_s$ is equal to or near 0. The statistical strength of the Spearman correlation has been demonstrated to be as robust as that of the parametric Pearson's $r$, especially for data sets with considerable range in values and reduced frequency of tied ranks.

In the Spearman rank correlation coefficient, the value of the 6 in the numerator stems from a conversion done to produce a scale with a similar interpretation to that of the Pearson's $r$. To understand the need for the conversion (as outlined by D. Griffiths), take into account the extreme cases of the rankings where the relationship is perfectly positively or negatively monotonic. If the rankings for each ordered pair ($N$) are identical, then the sum of the differences between the rankings for variable $x$ with ordered pair $y$ ($\Sigma d^2$) will be 0 and the value of $r_s$ will be 1. For example, the ranks of the pairs (1,1), (2,2), (3,3), (4,4), and (5,5) do not differ and illustrate a positive monotonic relationship with a perfectly positive correlation coefficient ($r_s = 1$). In contrast, if the rankings for each ordered pair ($N$) directly oppose one another, then the sum of the differences between the rankings for variable $x$ with ordered pair $y$ ($\Sigma d^2$) will be equal to $N\,(N^2 - 1)/3$. For example, the ranks of the pairs (1,5), (2,4), (3,3), (4,2), and (5,1) differ considerably and demonstrates a negative monotonic relationship, whereby $x$ and $y$ are inversely related to one another and $r_s$ will be more negative with increasing $N$. Consequently, the lower range of $r_s$ will be dependent on $N$ and the scale quantifying these associations will fluctuate. By finding an equation that passes through the extreme points in either direction from zero (i.e., (0, 1) and ($N$ $(N^2-1)/3$, $-1$) for this data set), the scale of the Spearman ranked correlation coefficient is adjusted to a standard scale. The resulting equation is $y = -6/N(N^2 - 1)\ x + 1$, where the slope of the line is $-6/N\ (N^2 - 1)x$ and the $y$ intercept is 1. This transformation produces a coefficient then that is not dependent on the number of ordered pairs ($N$) and varies similarly to the Pearson's $r$, between $-1$ and 1.

## Tied Rankings

As with other statistical techniques employing ordinal-level data, the Spearman rank correlation coefficient can be substantially influenced by the incidence of tied rankings. For example, a data set examining the relationship between annual income and vehicles owned would in all likelihood have a large number of tied ranks because the latter variable will tend to elicit a more limited response (most likely values would be between 0 and 3, with 1 and 2 vehicles owned being more common). As the number of tied ranks increases, the larger the difference between the sum of the square of ranks from the sum of the square of $N$ (the number of paired data values) and the larger the effect there will be on $r_s$. In these cases, a correction factor is needed to minimize the effect of tied ranks and produce an uninflated, more realistic representation of $r_s$. One common method used to correct for tied ranks is to sum the number of tied positions and divide by the number of tied scores, thus assigning each of the tied scores the mean rank value. For instance, a data set with the values of 1, 3, 6, 6, 6, 6, 7, and 9 would have the ranks of 1, 2, 3, 3, 3, 3, 7, and 8, respectively; thus, the tied values of 6 all receive an equal ranking of 3. Applying the correction factor, each of the tied rankings would instead be assigned a value of 4.5 (i.e., the sum of the tied positions is $3 + 4 + 5 + 6 = 18$ divided by the number of tied scores, 4). However, the effect of tied ranking on $r_s$ is only significant when the

fraction of tied ranking to the total number of values is sufficiently large, often a quarter or more of the entire data set. Computer statistical packages (e.g., IBM® SPSS® (PASW) 18.0, an IBM company) typically will automatically make these adjustments, if necessary.

## Statistical Significance

The Spearman correlation coefficient describes the strength of the association between the two ordinal variables, but it does not show whether that relationship is statistically significantly different from 0. Because the sample correlation coefficient ($r_s$) estimates the population correlation coefficient ($\rho_s$), a test statistic is needed to determine the degree of confidence that the relationship found from random sampled pairs is truly representative of the entire population. This step is particularly important when using relatively small sample sizes where there is an increased probability that a nonzero correlation coefficient results through chance sampling from an otherwise uncorrelated data set. The test for statistical significance examines whether the null hypothesis is confirmed or rejected in favor of an alternate hypothesis. The null hypothesis is that no relationship exists between the two variables $x$ and $y$ in the population data set and the correlation is zero ($H_0 : \rho_s = 0$). If the null hypothesis is true, then the two variables do not covary and are independent of one another. Conversely, if the null hypothesis is false, then the two variables are not mutually exclusive, and the alternative hypothesis that the two variables are significantly correlated within the population must be accepted. The alternate hypothesis ($H_a$) is either one of two forms, depending on the problem under consideration. For situations in which prior knowledge concerning the relationship between the two variables is unknown, a nondirectional or two-tailed significance test is applied, and the alternate hypothesis is that a relationship does exist between the two variables ($H_a : \rho_s \neq 0$). However, if the variables are anticipated to covary in one direction, then a one-tailed significance test is used and the alternate hypothesis is that the variables are related in either a positive ($H_a : \rho_s > 0$) or negative ($H_a : \rho_s < 0$) direction.

To test the significance of the Spearman rank correlation coefficient (or the Pearson's $r$), either a $Z$ or $t$ distribution is usually applied depending on the size of the sample population. Significance testing for larger sample sizes (usually when $N > \sim 30$) will use the $Z$ or normal distribution, a probability density function with a mean equal to zero, and a standard deviation equal to one. For the $Z$ distribution, the significant correlation test is governed by the value of the Spearman correlation coefficient and the number of degrees of freedom (i.e., the sample size minus one population variable being estimated, the population correlation coefficient) such that

$$Z_{r_s} = r_s\sqrt{n-1}.$$

For smaller sample sizes, the Student's $t$ distribution is more appropriate than the $Z$ distribution, because the deviations from the normal distribution increase with decreasing degrees of freedom. The $t$ distribution test statistic is also dependent on the strength of Spearman correlation coefficient and sample size and is commonly written as

$$t = r\sqrt{\frac{(n-2)}{(1-r^2)}}.$$

The larger the $Z$ or $t$ value, the less likely the relationship between the two variables is random and the alternate hypothesis should confidently be accepted in favor of the null hypothesis.

The $Z$ or $t$ scores obtained from significance testing can then be converted to a $p$ value using a probability table that signifies the proportional area under a normal or Student's $t$ distribution curves, respectively. The probability value or $p$ value represents the probability of rejecting the null hypothesis that no relationship exists between the variables when in fact, it is true (i.e., a Type I error). Large $Z$ and $t$ scores generally yield progressively smaller $p$ values and, thus, decrease the likelihood of making a Type I error. For example, suppose the relationship between two variables produce a Spearman rank correlation coefficient of 0.50 based on a sample size of 22. Because the sample size is relatively small, a significance test statistic and probability table based on the Student's $t$ distribution is used. The computation produces a $t$ equal to 2.6

and an associated $p$ value of .017, suggesting a small probability that the null hypothesis is true ($H_0 : \rho_s = 0$) and that the two variables are most likely significantly correlated. Large Spearman rank correlation coefficients ($> .50$) in conjunction with larger sample sizes will generate $p$ values closer to 0 and indicate a statistically significant relationship; however, the researcher must determine whether that relationship is meaningful and justified.

*Jill S. M. Coleman*

***See also*** Coefficients of Correlation, Alienation, and Determination; Correlation; Distribution; Nonparametric Statistics; Normality Assumption; Ordinal Scale; Pearson Product-Moment Correlation Coefficient; $p$ Value; Significance Level, Concept of; SPSS; Student's $t$ Test

### Further Readings

Conover, W. J. (1999). *Practical non-parametric statistics.* (3rd ed.). New York: Wiley.

Gibbons, J. D. (1993). *Non-parametric measures of association.* Thousand Oaks, CA: Sage.

Griffiths, D. (1980). A pragmatic approach to Spearman's rank correlation coefficient. *Teaching Statistics, 2,* 10–13.

McGrew, J. C., Jr., & Monroe, C. B. (1999). *An introduction to statistical problem solving in geography* (2nd ed.). New York: McGraw-Hill.

Spearman, C. E. (1904). Proof and measurement of association between two things. *American Journal of Psychology, 15,* 72–101.

# SPECIFICITY

The accuracy of a test in making a dichotomous classification might be evaluated by one of four related statistics. Of these four statistics, specificity is used to evaluate the probability of correctly identifying the absence of some condition or disease state. For example, specificity might be used in a medico-legal setting to describe that a particular test has 95% probability of detecting that a head-injured patient is not malingering. Specificity is calculated as the proportion of true negative cases divided by all cases without the condition.

## Calculating Specificity Scores

Specificity is calculated based on the relationship of the following two types of dichotomous outcomes: (1) the true state of affairs and (2) the outcome of the test or collection of tests. The true state of affairs is known either via experimental assignment of some condition or through classification based on some gold standard test. The outcome of the test is typically referred to as being positive (indicating the condition is present) or negative (the condition is not present). Based on these two types of dichotomous outcomes, there are four possible outcome variables, which are defined as follows:

*True negative* = the number of cases with a negative test outcome that *do not* have the condition

*True positive* = the number of cases with a positive test outcome that *do* have the condition

*False negative* = the number of cases with a negative test outcome that *do* have the condition (Type II error)

*False positive* = the number of cases with a positive test outcome that *do not* have the condition (Type I error)

Specificity is computed from the pool of individuals who do not have a particular condition or disease. Specificity is the number of true negative cases divided by the number of true negatives plus the number of false positives. This is distinct from sensitivity, which is computed as the number of true positive cases divided by the number of true positives plus false negatives.

The specificity of a test is typically not fixed but might vary depending on the cutoff used to define a negative test outcome. To demonstrate this, consider an example in which cognitive performance was measured from 100 head injury cases, 50 of which are instructed to perform to the best of their true ability level on measures of cognition (symptom validity test) and the other 50 are instructed to perform poorly (malinger). Table 1 shows, from left to right, simulated data showing cognitive performance levels and the corresponding number of patients scoring at that performance level who were instructed to perform their best. If a negative test is defined as being any patient with

**Table 1** Example Showing Performance Levels Among Head Injury Patients Instructed to Perform to the Best of Their Ability

| *Symptom Validity Test Score* | *Number of Patients* |
|---|---|
| 20 | 5 |
| 30 | 10 |
| 40 | 15 |
| 50 | 20 |

a performance level of 40 or higher on the symptom validity test, then 35 of the 50 patients instructed to perform their best would have a negative test outcome (sample size from the groups with scores of 40 or 50). Given this cutoff, then, the specificity of the procedure would be calculated as: [35 (true negatives) ÷ [35 (true negatives) + 15 (false positives)]] = 70%. In contrast, if a negative test is defined as a performance level of 30 or higher, then 45 of the 50 patients instructed to perform their best would have a negative test outcome (resulting in a specificity of 90% [45 ÷ [45 + 5]]).

The patients instructed to perform poorly are not listed in Table 1 because they would by definition either be true positive or false negative cases, which are not part of the specificity calculation. They would instead be pertinent to the calculation of sensitivity. This example demonstrates how specificity and sensitivity are independent measures because they are calculated on two different samples of individuals (those without or with a particular condition).

### Specificity in Research Design

Specificity is typically used to demonstrate or evaluate the accuracy of a test for correctly ruling out the presence of some condition or disease state. This measure of a test's accuracy in classification is especially important in settings where a false positive is extremely costly. Consider the example of head-injured patients and malingering. If a patient is classified on the basis of some test as malingering, then this information might be used in the medico-legal setting to deny that patient financial compensation for his or her condition and thereby limit his or her ability to obtain services for treatment. In addition, there is a significant social stigma in being assigned such a classification. Because of these costs, it is imperative that the number of false positives be minimized, thereby maximizing specificity.

Specificity is not the only relevant measure of test accuracy. Although specificity provides information about the accuracy of a procedure for ruling out the presence of some condition, it does nothing to indicate that test's ability to identify a condition accurately. The accuracy in making a positive test outcome would be defined by sensitivity, which is described in a separate entry of this encyclopedia. Also, specificity is not the most pertinent measure of test accuracy in clinical settings where decisions are being made about an individual's test outcome and the true diagnoses or classification is unknown. In this setting, what is of greater interest is the confidence in interpreting an individual test outcome, which is defined by positive and negative predictive values. Unlike specificity, predictive values take into account the population base-rate of the phenomenon of interest.

*Charles W. Mathias*

***See also*** Dichotomous Variable; False Positive; Sensitivity; Type I Error; Type II Error

**Further Readings**

Bianchini, K. J., Mathias, C. W., & Greve, K. W. (2001). Symptom validity testing: A critical review. *The Clinical Neuropsychologist, 15,* 19–45.

Glaros, A. G., & Kline, R. B. (1988). Understanding the accuracy of tests with cutting scores: The sensitivity, specificity, and predictive value model. *Journal of Clinical Psychology, 44,* 1013–1023.

Loong, T.-W. (2003). Understanding sensitivity and specificity with the right side of the brain. *British Medical Journal, 327,* 716–719.

## SPHERICITY

Sphericity is an assumed characteristic of data analyzed in repeated measures analysis of variance (ANOVA). Sphericity refers to the equality of variances of the differences between treatment conditions. Violations of the sphericity assumption do not invalidate a repeated measures ANOVA but do necessitate using corrected *F* ratios.

To illustrate sphericity, imagine you were conducting a study administering placebo, a low dose of alcohol, and a high dose of alcohol to young adults and measuring subjective ratings of alcohol intoxication. You wish to analyze your results using a repeated measures ANOVA, but to do so you must check whether the sphericity assumption is violated. Sphericity can be measured by the following:

1. Comparing each pair of treatment levels
2. Calculating the differences between each pair of scores
3. Calculating the variances of these differences
4. Determining whether the variances of the different treatment pairs are significantly different

For the sphericity assumption to hold up, the variances of the different treatment pairs must be about equal, illustrated in the following:

$$\text{Variance}_{\text{placebo/low dose}} \approx \text{Variance}_{\text{placebo/high dose}} \approx \text{Variance}_{\text{low dose/high dose}}$$

## Assessing Departures From Sphericity With Mauchly's Test

Mauchly's test is a statistical procedure that might be used to determine whether the sphericity assumption has been met for a given data set. Repeated-measures ANOVA performed with the IBM® SPSS® (PASW) 18.0 statistical program (an IBM company) will automatically include Mauchley's test. If the Mauchley's test statistic is not significant (i.e., $p > .05$), then it might be assumed that the variances of the differences between treatment conditions are roughly equal, and the sphericity assumption has been met. However, if the Mauchley's test statistic is significant (i.e., $p < .5$), then the variances are not equal and the assumption of sphericity has been violated.

Some caution is in order on using Mauchley's test to check for violations of sphericity. This test might be underpowered with small sample sizes, which might lead to violations of the sphericity assumption going undetected. Furthermore, Mauchley's test does not provide information regarding the degree to which sphericity has been violated.

## Correcting for Violations of Sphericity

There are corrections that might be used when the sphericity assumption has been violated, although they result in the loss of some statistical power. The most commonly used are the Greenhouse–Geisser correction (developed by Samuel Greenhouse and Seymour Geisser), the Huynh-Feldt correction (developed by Huynh Huynh and Leonard Feldt), and a hybrid of the two. All three corrections decrease the degrees of freedom and thereby increase the $p$ value needed to obtain significance.

These corrections first require generating a value known as epsilon. Epsilon varies between $1/(k - 1)$ ($k$ = number of repeated treatments) and 1. The closer epsilon is to 1.0, the closer data are to meeting the sphericity assumption. The lower limit of epsilon naturally varies based on the number of repeated treatments. For example:

$$\text{When } k = 3, \text{ the lower limit of epsilon} = 1/(3-1) = 0.5.$$

$$\text{When } k = 5, \text{ the lower limit of epsilon} = 1/(5-1) = 0.25.$$

The Greenhouse–Geisser correction generates a conservative estimate of epsilon and thus might generate an overly conservative correction. The Huynh-Feldt correction generates a more liberal estimate of epsilon, and this correction might be more appropriate when the Greenhouse–Geisser estimate of epsilon is greater than 0.75. However, it has also been argued that the Huynh-Feldt estimate of epsilon is too liberal (overestimates sphericity), and thus it might be more appropriate to average the epsilon values obtained from both and use that averaged value for adjusting degrees of freedom.

Another alternative is to use a multivariate ANOVA (MANOVA) for repeated measures testing, as the MANOVA is not dependent on the sphericity assumption. The MANOVA is less powerful than the repeated measures ANOVA, but it might be preferable when study sample sizes are large and epsilon values are low.

*Ashley Acheson*

*See also* Analysis of Variance (ANOVA); Greenhouse–Geisser Correction; Mauchly Test; Multivariate Analysis of Variance (MANOVA); Variance

### Further Readings

Field, A. (1998). A bluffer's guide to . . . sphericity. *The British Psychological Society: Mathematical, Statistical & Computing Section Newsletter, 6,* 13–22.

Field, A. (2005). *Discovering statistics using SPSS* (2nd ed.). Thousand Oaks, CA: Sage.

Greenhouse, S. W., & Geisser, S. (1959). On the methods in the analysis of the profile data. *Psychometrika, 24,* 95–112.

Huynh, H., & Feldt, L. S. (1976). Estimation of the Box correction for degrees of freedom from sample data in randomised block and split-plot designs. *Journal of Educational Statistics, 1,* 69–82.

## Split-Half Reliability

Measurement is fundamental to almost all forms of research and applied science. To conduct quantitative research, scientists must measure at least one variable. For example, researchers studying the effect of social rejection on self-esteem must measure participants' self-esteem in some way. Similarly, to apply scientific knowledge, practitioners often rely heavily on measurement. For example, school psychologists measure children's academic and cognitive aptitudes to place them in appropriate classes and to identify potential academic difficulties. Given the importance of measurement, researchers and practitioners must evaluate the quality of the measurement tools that they use. Reliability is a key facet of measurement quality, and split-half reliability is a method of estimating the reliability of a measurement instrument.

### Reliability

Briefly stated, reliability reflects the precision of scores obtained from a measurement instrument—how closely participants' scores on the instrument correspond to their real characteristics. Unfortunately, many factors can interfere with measurement in any scientific domain, some of which are unsystematic sources of measurement error. Such factors artificially inflate some participants' scores and deflate others' scores in a random, or unsystematic, way. In behavioral research, these factors can include guessing, poorly written items, fatigue, misreading test items, and temporary mood states.

Consider, for example, a participant in a study involving a measure of trait self-esteem (i.e., the degree to which a person sees himself or herself in a generally positive way). Imagine that the participant actually has a high level of trait self-esteem, generally having a positive view of himself or herself. Unfortunately, one or two of the self-esteem questionnaire's items are worded in a confusing manner (e.g., "I rarely feel as if I don't have low self-esteem"). Such items can elicit confused responses that do not reflect accurately the person's truly high level of self-esteem, thereby introducing error and imprecision into the measurement process. As an index of measurement precision, reliability reflects the degree to which test scores are free of unsystematic measurement error.

Reliability cannot be known directly, so it must be estimated. Much as a person's self-esteem is not directly observable and must be estimated from his or her test scores, reliability is not directly observable and must be estimated from a set of test scores. As a fundamental facet of reliability, measurement error cannot be known in reality—researchers cannot truly know the degree to which a respondent's scores are affected by fatigue, confusing wording, mood states, or any of the many factors potentially affecting test scores. Consequently, reliability must be estimated from the scores obtained on the measurement instrument itself. Split-half reliability is one of many approaches to estimating the reliability of scores on a measurement instrument.

### Computing and Interpreting Split-Half Reliability

The split-half method of estimating reliability is most directly applicable to instruments that have multiple items. Indeed, many instruments in behavioral research are tests, questionnaires,

**Table 1** Split-Half Reliability Example Data

| | *Item* | | | | | *Split 1* | | *Split 2* | |
|---|---|---|---|---|---|---|---|---|---|
| *Person* | 1 | 2 | 3 | 4 | *Total Score* | 1 & 3 | 2 & 4 | 1 & 4 | 2 & 3 |
| 1 | 3 | 5 | 3 | 3 | 14 | 6 | 8 | 6 | 8 |
| 2 | 6 | 3 | 5 | 5 | 19 | 11 | 8 | 11 | 8 |
| 3 | 5 | 6 | 2 | 2 | 15 | 7 | 8 | 7 | 8 |
| 4 | 2 | 2 | 3 | 2 | 9 | 5 | 4 | 4 | 5 |
| Mean | 4.00 | 4.00 | 3.25 | 3.00 | 14.25 | 7.25 | 7.00 | 7.00 | 7.25 |
| *SD* | 1.58 | 1.58 | 1.09 | 1.22 | 3.56 | 2.28 | 1.73 | 2.55 | 1.30 |

*Note*: *SD* = standard deviation.

inventories, or surveys that include two or more items.

Consider the hypothetical set of responses in Table 1. Imagine that a researcher wishes to estimate the reliability of a four-item test of trait self-esteem, in which each item presents a statement relevant to self-esteem (e.g., "I often feel that I am a good person"). People respond to each item using a seven-point scale indicating their level of agreement with the statements (e.g., 1 = strongly disagree, 4 = neutral, and 7 = strongly agree)—thus, larger numbers reflect greater self-esteem. Peoples' responses are summed to create a total score indicating their level of trait self-esteem. Of course, many good tests include negatively keyed items, for which an endorsement or agreement reflects a low level of the characteristic being measures (e.g., "I rarely feel like I'm a good person"). Such items must be reverse scored before scoring the scale and evaluating reliability. As shown in Table 1, Person 2 has the highest level of self-esteem and Person 4 has the lowest. Being aware that scores on the instrument might be affected by measurement error, the researcher estimates the reliability of these scores.

The split-half method of estimating reliability can be viewed as a three-step process. First, the test is divided into two subtests, and participants' scores are computed for each subtest. For example, the "split 1" columns in Table 1 present peoples' scores for two subtests—one subtest formed by summing the odd items and one formed by summing the even items (ignore the "split 2" columns for the moment).

Second, the Pearson correlation between the two subtests is computed. This "split-half correlation" reflects the consistency between the two parts of the test. In this regard, note the discrepancy between the two subtests—the differences among peoples' scores on the first subtest are somewhat inconsistent with the differences among their scores on the second. Specifically, on the first subtest, person 2 has the highest score, followed by person 3, and then person 1; however, on the second subtest, persons 2, 3, and 1 have identical scores. That is, subtest 1 indicates differences in these three peoples' self-esteem, but subtest 2 indicates no differences in their self-esteem. This inconsistency between the subtests suggests an imperfect test; a very strong test should include subtests that were strongly consistent with each other. Indeed, the correlation between the two halves is .57 (recall that correlations can be as large as 1.0, indicating perfect consistency). Although this correlation is not a perfect 1.0, it is large enough to indicate a fair degree of consistency—in both subtests, person 4's score is smaller than the other peoples' scores. Because it arises from consistency between parts of a test, split-half reliability is an "internal consistency" approach to estimating reliability.

Third, the split-half correlation is entered into an equation producing the estimate of reliability. The equation is a form of the Spearman–Brown prophecy formula, tailored to the spit-half approach:

$$\text{Split-half estimate} = \frac{2r_{SH}}{1 + r_{SH}},$$

where $r_{SH}$ is the split-half correlation obtained in the previous step. Thus, for the previous example, the split-half estimate of reliability is .73:

$$\text{Split-half estimate} = \frac{2(.57)}{1 + .57} = \frac{1.14}{1.57} = .73.$$

This result is an estimate of the reliability of the test scores, and it provides some support for the quality of the test scores. In theory, reliability ranges from zero to 1.0, with higher values reflecting better reliability. In psychological measurement, experts cite reliability of .70 or .80 as being acceptable for research purposes, with even higher reliability being desirable when using test scores for decisions about individuals (e.g., when assigning students to classes on the basis of aptitude scores). Based on these guidelines, the example's test scores are reasonably reliable for research purposes.

A split-half estimate of reliability can be obtained from some statistical software packages. For example, if item-level responses are entered into IBM® SPSS® (PASW) 18.0, an IBM product, then the "Scale . . . Reliability Analysis" procedure offers the "split-half" model as an option. SPSS automatically creates subtests by grouping the first half of the items together and the second half of the items together (note that this is not an ideal method of creating subtests).

## Problems and Contemporary Practice

Although split-half reliability was important in the evolution of measurement theory, it is now reported infrequently. Indeed, during the past few decades, the use of split-half reliability has been decreasing while the use of other methods (e.g., coefficient alpha) has been increasing. This trend arises from at least three problems with the split-half method of estimating reliability.

First, subtests can be formed in many ways that can produce dramatically different estimates of reliability. Subtests can be formed not only by combining odd items and combining even items (as was done for the previous example), but also by first-half and second-half (as is done by SPSS), by random assignment of items to halves, and by various other methods. For example, the "split 2" columns in Table 1 present subtest scores for a different method of creating subtests. Specifically, the first item was combined with the fourth item, and the second item was combined with the third item. The correlation between these subtests is .68, which produces an estimated reliability of .81; this estimate is larger than the estimate obtained from the previous split. In contrast, combining item 1 and 2, and combining items 3 and 4 (as is done by SPSS) produces a split-half correlation of only .09 and an estimated reliability of only .16—dramatically lower than the other estimates. Because the many ways of splitting a test can produce varying estimates of reliability, there is no single value that can be taken as *the* split-half estimate of reliability.

A second problem with spilt-half reliability is that it does not account for measurement error unique to a single testing occasion. As an internal consistency approach to estimating reliability, the split-half method reflects the item-by-item effects of random measurement error (e.g., items that are confusing or do not fit with the rest of the items, the effect of fluctuating attention, and guessing on some items but not others); however, it does not reflect error affecting responses to the entire set of items. For example, it does not reflect the possibility that a respondent's temporary sad mood artificially deflates her score across all the items on a measure of trait self-esteem. That is, temporary mood states might affect a person's responses to the entire test, but internal consistency estimates such as split-half reliability will not reflect the effects of this form of measurement error.

A third problem is that the accuracy of split-half reliability as a method for estimating reliability depends on several assumptions that are unlikely to be valid in research applications (e.g., the assumption that subtests have identical standard deviations). Because the underlying assumptions might not be true, the split-half approach might not provide good estimates of reliability in many research contexts. Note, however, that all methods of estimating reliability rest on certain assumptions; this is not unique to split-half reliability.

Considering the problems associated with the split-half method as a way of estimating reliability, contemporary researchers tend to use different methods. Specifically, researchers are much more likely to use coefficient alpha (also called Cronbach's alpha), which avoids some problems associated with the split-half approach (e.g., the problem of splitting a test into halves) and is available through statistical software such as SPSS and SAS. Nevertheless, split-half reliability has played

an important role in the evolution of reliability theory and in applied measurement.

*R. Michael Furr*

*See also* Coefficient Alpha; Internal Consistency Reliability; Pearson Product-Moment Correlation Coefficient; Psychometrics; Reliability; Spearman–Brown Prophecy Formula

### Further Readings

Charter, R. A. (2003). A breakdown of reliability coefficients by test type and reliability method, and the clinical implications of low reliability. *Journal of General Psychology, 130,* 290–304.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 105–146). Washington, DC: American Council on Education.

Furr, R. M., & Bacharach, V. R. (2008). *Psychometrics: An introduction.* Thousand Oaks, CA: Sage.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences.* San Francisco: W. H. Freeman.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory.* New York: McGraw-Hill.

Schmidt, F. L., & Hunter, J. E. (1999). Theory testing and measurement error. *Intelligence, 27,* 183–198.

# Split-Plot Factorial Design

It is often inconvenient, costly, or even impossible to perform a factorial design in a completely randomized fashion. An alternative to a completely randomized design is a split-plot design. The use of split-plot designs started in agricultural experimentation, where experiments were carried out on different plots of land. Classical agricultural split-plot experimental designs were full factorial designs but run in a specific format. The key feature of split-plot designs is that levels of one or more factors are assigned to entire plots of land referred to as whole plots or main plots, whereas levels of other factors are assigned to parts of these whole or main plots. These parts are called subplots or split-plots. Split-plot designs thus have two types of experimental units, whole plots and subplots. The smaller experimental units, the subplots, are nested within the larger ones, the whole plots.

## A Sample Design

Consider an experiment in which the effect of different types of nitrogen and grain varieties on the grain yield is investigated. Usually, it is convenient to treat entire plots of land (the whole plots) with the same type of nitrogen, whereas different grain varieties can typically be planted on small plots of land (the subplots) obtained by splitting each whole plot that was treated with one kind of nitrogen only. In this experimental setup, the first factor, the type of nitrogen, is called the whole-plot factor because its levels are applied to whole plots. The factor grain variety is called the subplot factor of the experiment because its levels are applied to the subplots. A schematic representation of a split-plot design involving type of nitrogen as the whole-plot factor and grain variety as the subplot factor is displayed in Figure 1. The four levels of the whole-plot factor are denoted by $N_1$–$N_4$, whereas the three levels of the subplot factor are

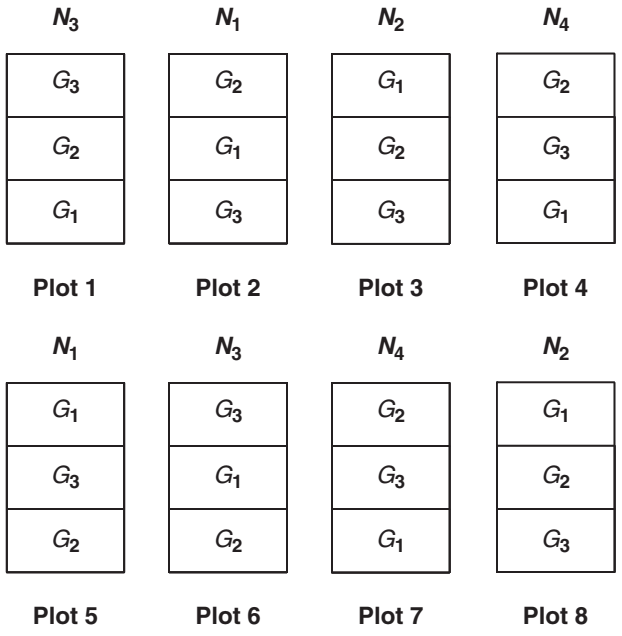



**Figure 1** Classical Agricultural Factorial Split-Plot Design Where $N_1$–$N_4$ Represent the Levels of the Type of Nitrogen and $G_1$–$G_3$ Denote the Levels of the Factor Grain Variety

*Note:* The design has eight whole plots, each of which is split into three subplots.

represented by $G_1-G_3$. The split-plot design in this example has only one whole-plot factor and one subplot factor. Nevertheless, split-plot factorial designs with more than one whole-plot factor and more than one subplot factor can be constructed too. An example of a design with two whole-plot factors is given in the following discussion.

## Design Layout

The layout of a split-plot design resembles that of a randomized block design. The key difference between split-plot designs and randomized block designs is that, in randomized block designs, the factor level combinations are randomly assigned to the experimental units in the blocks. In split-plot designs, a completely random assignment is impossible because all the subplots within one whole plot must be treated with the same level of the whole-plot factor(s). Because of this restriction on the assignment of factor-level combinations to the plots, only a restricted form of randomization is possible. The restricted randomization is carried out in two steps. First, the whole-plot factor-level combinations are randomly assigned to the whole plots. Next, the subplot factor-level combinations are randomly assigned to the subplots within each whole plot. This is done so that each subplot factor-level combination appears once in every whole plot.

## Analysis

In the analysis of data from split-plot designs, the whole plots act as the levels of a blocking factor. Therefore, along with the effects of the whole-plot and subplot factors, a random effect is included in the statistical model for each of the whole plots, and advantages similar to the ones for block designs can be obtained using split-plot designs. For example, the effect of the grain varieties in Figure 1 is obtained from the differences between the yields within each whole plot. These differences eliminate the whole-plot effect, so that the effect of the grain varieties is measured with greater precision than the effect of the type of nitrogen. The calculation of the effect of the nitrogen type is not as precise because it involves computing and comparing average yields over all the subplots within a whole plot, and errors between subplots in different whole plots are usually larger than errors between subplots in a single whole plot. For this reason, the effect of the whole-plot factor, type of nitrogen, is tested against the whole-plot error, and the effect of the subplot factor, grain variety, is tested against the subplot error. An attractive feature of split-plot designs is that interaction effects between a whole-plot factor and a subplot factor can be estimated with the same precision as the subplot factor effects, so that they are also tested against the subplot error. Thus, a split-plot design is an excellent design option if powerful inference is desired regarding the subplot factors and the interactions between whole-plot factors and subplot factors.

In a classic split-plot analysis, the whole-plot and subplot factors are treated as categorical, and the analysis of variance (ANOVA) model used to analyze the split-plot data can be written as

$$Y_{ijk} = \mu + \delta_{i(j)} + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \varepsilon_{ijk},$$
$$i = 1, \ldots, w; j = 1, \ldots, a; k = 1, \ldots, b,$$

where $Y_{ijk}$ is the response for the $j$th type of nitrogen and the $k$th grain variety in whole plot $i$, $\mu$ is the overall mean, $\delta_{i(j)}$ is the random effect of the $i$th whole plot nested within the $j$th level of the whole-plot factor, type of nitrogen, $\alpha_j$ is the main effect of the $j$th type of nitrogen, $\beta_k$ is the main effect of the $k$th grain variety, $(\alpha\beta)_{jk}$ is the interaction effect of the $j$th type of nitrogen and the $k$th grain variety, $\varepsilon_{ijk}$ is the random error, $a$ denotes the number of whole-plot factor levels, $b$ is the number of subplot factor levels, and $w$ is the number of whole plots at each whole-plot factor level. In this example, $a$, $b$, and $w$ equal 4, 3, and 2, respectively. To make the split-plot model estimable, the factor effects $\alpha_j$, $\beta_k$, and $(\alpha\beta)_{jk}$ are subject to the constraints

$$\sum_{j=1}^{a} \alpha_j = 0, \ \sum_{k=1}^{b} \beta_k = 0,$$
$$\sum_{j=1}^{a} (\alpha\beta)_{jk} = 0 \text{ for all } k,$$

and

$$\sum_{k=1}^{b} (\alpha\beta)_{jk} = 0 \text{ for all } j.$$

The effects $\alpha_j$, $\beta_k$, and $(\alpha\beta)_{jk}$ are fixed effects. They are referred to as whole-plot factor effects, subplot factor effects, and whole-plot-by-subplot interaction effects, respectively. The remaining effects in the model, $\delta_{i(j)}$ ($i = 1, \ldots, w$; $j = 1, \ldots, a$) and $\varepsilon_{ijk}$ ($i = 1, \ldots, w$; $j = 1, \ldots, a$; $k = 1, \ldots, b$), are random effects. They are called the whole-plot errors and subplot errors, respectively. The whole-plot errors $\delta_{i(j)}$ are assumed to be independently normally distributed with zero mean and variance $\sigma_\delta^2$. The subplot errors $\varepsilon_{ijk}$ are assumed to be independently normally distributed with zero mean and variance $\sigma_\varepsilon^2$. It is also assumed that each of the whole-plot errors is independent from each of the subplot errors. Under these assumptions, responses from observations in different whole plots are independent, whereas responses from observations within the same whole plot are correlated. The assumed correlation structure is special in the sense that, although two observations in a given whole plot are correlated in advance of the experiment, they are independent once a whole plot has been selected. This correlation structure is usually referred to as a compound symmetric structure.

A general ANOVA table for a classic split-plot factorial design with one whole-plot factor and one subplot factor is displayed in Table 1. In the table, the whole-plot and subplot factors are denoted by $A$ and $B$, respectively, and $a$, $b$, and $w$ denote the number of levels of factor $A$, the number of levels of factor $B$, and the number of whole plots at each whole-plot factor level, respectively. The number of observations in the split-plot design equals $N = abw$, whereas the total number of whole plots equals $q = aw$.

A key characteristic of split-plot designs is that the whole-plot factor effects, $\alpha_j$, are estimated less precisely than the subplot factor effects, $\beta_k$. This implies that split-plot designs result in a small power for detecting whole-plot effects and a high power for detecting subplot effects. Because the interaction effects between a whole-plot factor and a subplot factor, $(\alpha\beta)_{jk}$, are estimated precisely as well, split-plot designs also have a high power for detecting these. Ignoring the split-plot nature of the experiment and analyzing the data from a split-plot factorial design as if they were data from a completely randomized factorial design leads to an underestimation of the variance of the whole-plot factor effect estimates and to an overestimation of the variance of the subplot factor effect estimates and the whole-plot-by-subplot interaction effect estimates. Therefore, ignoring the split-plot nature of the design results in declaring whole-plot effects to be significant too often and declaring subplot effects and the whole-plot-by-subplot interaction effects to be significant not often enough.

The data for the split-plot design depicted in Figure 1 are shown in Table 2. The table clearly shows that the same type of nitrogen was used for every observation in a given whole plot, whereas every grain variety was used exactly once in every whole plot. The corresponding ANOVA table is displayed in Table 3. In the ANOVA table, the effect of the whole-plot factor, type of nitrogen, is tested against the whole-plot error. The effect of the subplot factor, grain variety, is tested against the subplot error, as is the interaction effect between the type of nitrogen and grain variety. The main effect of the subplot factor is highly significant, as well as the interaction effect. The main effect of the whole-plot factor is not significant.

## Application

Split-plot factorial designs are applied not only in agriculture but also in many other application areas. Obviously, the whole plots and the subplots then no longer correspond to plots of land. For example, one experimental factor in many industrial experiments is an oven temperature. Typically, an oven has enough space to accommodate several experimental units, to which other factors are applied. In such experiments, each run of the oven plays the role of a whole plot, whereas the different positions in the oven are similar to the subplots in classic agricultural split-plot designs. Another typical situation where split-plot designs are used is in prototype experimentation. Consider for example a wind tunnel experiment where the factors under investigation are the front ride height of a car (studied at two levels), the car's rear ride height (studied at two levels as well), and the driving condition (studied at three levels). A duplicated, completely randomized factorial design would require building 24 different prototypes and putting a new prototype in the wind tunnel for every single experimental run. Using the split-plot

**Table 1** ANOVA Table for a Classic Split-Plot Factorial Design With One Whole-Plot Factor, A, and One Subplot Factor, B

| *Source of variation* | *SS* | *df* | *MS* | *E(MS)* |
|---|---|---|---|---|
| Whole plots | | | | |
| Factor *A* | $SSA = bw\sum_{j=1}^{a}(\bar{Y}_{.j.} - \bar{Y}_{...})^2$ | $a - 1$ | *MSA* | $\sigma_\varepsilon^2 + b\sigma_b^2 + bw\frac{\sum_{j=1}^{a}\alpha_j^\varepsilon}{a-1}$ |
| Whole-plot error | $SSW(A) = b\sum_{i=1}^{w}\sum_{j=1}^{a}(\bar{Y}_{ij.} - \bar{Y}_{.j.})^2$ | $a(w - 1)$ | *MSW(A)* | $\sigma_\varepsilon^2 + b\sigma_b^2$ |
| Subplots | | | | |
| Factor *B* | $SSB = aw\sum_{k=1}^{b}(\bar{Y}_{..k} - \bar{Y}_{...})^2$ | $b - 1$ | *MSB* | $\sigma_\varepsilon^2 + aw\frac{\sum_{k=1}^{b}\beta_k^2}{b-1}$ |
| *AB* interaction | $SSAB = w\sum_{j=1}^{a}\sum_{k=1}^{b}(Y_{.jk} - \bar{Y}_{.j.} - \bar{Y}_{..k} + \bar{Y}...)^2$ | $(a - 1)(b - 1)$ | *MSAB* | $\sigma_\varepsilon^2 + w\frac{\sum_{j=1}^{a}\sum_{k=1}^{b}(\alpha\beta)_{jk}^2}{(a-1)(b-1)}$ |
| Subplot error | $SSB.W(A) = \sum_{i=1}^{w}\sum_{j=1}^{a}\sum_{k=1}^{b}(Y_{ijk} - \bar{Y}_{ij.} - \bar{Y}_{.jk} + \bar{Y}_{.j.})^2$ | $a(w - 1)(b - 1)$ | *MSB.W(A)* | $\sigma_\varepsilon^2$ |
| Total | $SSTO = \sum_{i=1}^{w}\sum_{j=1}^{a}\sum_{k=1}^{b}(Y_{ijk} - \bar{Y}_{...})^2$ | $abw - 1$ | | |

**Table 2** Data for the Agricultural Split-Plot Factorial Design Depicted in Figure 1

| *WP* | *N* | *G* | *y* | *WP* | *N* | *G* | *y* | *WP* | *N* | *G* | *y* | *WP* | *N* | *G* | *y* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 3605 | 3 | 3 | 1 | 4368 | 5 | 1 | 1 | 3100 | 7 | 3 | 1 | 5066 |
| 1 | 1 | 2 | 3082 | 3 | 3 | 2 | 4116 | 5 | 1 | 2 | 3492 | 7 | 3 | 2 | 4213 |
| 1 | 1 | 3 | 3281 | 3 | 3 | 3 | 570 | 5 | 1 | 3 | 2569 | 7 | 3 | 3 | 1260 |
| 2 | 2 | 1 | 4097 | 4 | 4 | 1 | 5300 | 6 | 2 | 1 | 4762 | 8 | 4 | 1 | 5432 |
| 2 | 2 | 2 | 3562 | 4 | 4 | 2 | 3419 | 6 | 2 | 2 | 3916 | 8 | 4 | 2 | 4257 |
| 2 | 2 | 3 | 1644 | 4 | 4 | 3 | 1300 | 6 | 2 | 3 | 3335 | 8 | 4 | 3 | 664 |

*Notes:* The whole plots of the study are indicated in the column labeled *WP*. The columns labelled *N*, *G*, and *y* contain the type of nitrogen, the grain variety, and the observed responses, respectively.

**Table 3** ANOVA Table for the Split-Plot Data in Table 2

| *Source of variation* | *SS* | *df* | *MS* | *F* | *p* |
|---|---|---|---|---|---|
| Whole plots | | | | | |
| Type of nitrogen | 458,750.83 | 3 | 152,916.94 | 0.36 | .7888 |
| Whole-plot error | 1,718,688.33 | 4 | 429,672.08 | | |
| Subplots | | | | | |
| Grain variety | 29,829,030.58 | 2 | 14,914,515.29 | 79.11 | < .0001 |
| Nitrogen × grain | 10,937,914.42 | 6 | 1,822,985.74 | 9.67 | .0026 |
| Subplot error | 1,508,295.67 | 8 | 188,536.96 | | |
| Total | 44,452,679.83 | 23 | | | |

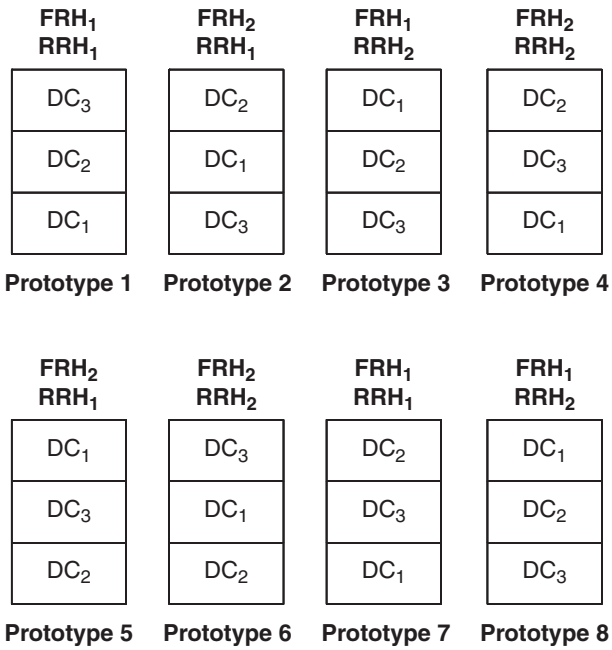



**Figure 2** Classic Industrial Split-Plot Design

*Notes:* $FRH_1$ and $FRH_2$, $RRH_1$ and $RRH_2$, and $DC_1$–$DC_3$ represent the levels of the factors' front ride height, rear ride height, and driving condition, respectively. The design has eight whole plots, each of which contains three observations and corresponds to a prototype.

design with 24 observations schematically visualized in Figure 2 would require only eight prototypes to be built. Each of the eight prototypes would then be tested under each of the three driving conditions successively. The two levels of the front ride height are denoted by $FRH_1$ and $FRH_2$. The two levels of the rear ride height are denoted by $RRH_1$ and $RRH_2$. The three driving conditions are denoted by $DC_1$–$DC_3$. In this design, the prototypes serve as the whole plots of the design, and the front ride height and the rear ride height are the whole-plot factors. Finally, split-plot designs also are often used in robust parameter design, where the researcher is interested in the precise estimation of interactions between control factors and noise factors. From these applications, it is clear that a more general name would be appropriate for split-plot designs. Several authors have used the term split-unit design as an alternative to split-plot design.

In some settings, for instance those involving large numbers of experimental factors, split-plot factorial designs will not be feasible. In such cases, fractional factorial split-plot designs or response surface designs run in a split-plot format can be considered. Industrial fractional factorial split-plot

designs with few whole plots and runs are often analyzed graphically using normal probability plots or half-normal probability plots. The analysis of data from response surface split-plot designs is done using a generalized least squares regression approach.

Experimental settings also exist where split-split-plot designs or strip-plot designs (which are also called strip-block or criss-cross designs) are more convenient than split-plot designs. Both split-split-plot designs and strip-plot designs are natural extensions of split-plot designs, which were also originally used in agricultural research but can be useful in other application areas too. Split-plot, split-split-plot, and strip-plot designs belong to the family of multistratum designs.

*Peter Goos*

***See also*** Analysis of Variance (ANOVA); Block Design; Factorial Design; Response Surface Design

**Further Readings**

Federer, W. T., & King, F. (2007). *Variations on split plot and split block experiment designs*. New York: Wiley.
Goos, P. (2002). *The optimal design of blocked and split-plot experiments*. New York: Springer.
Hinkelmann, K., & Kempthorne, O. (2008). *Design and analysis of experiments*. New York: Wiley.

# SPSS

IBM® SPSS® Statistics Version 18.0, formerly called PASW Statistics and most commonly referred to as simply SPSS, is a computer-based data management and inferential statistical analysis program. It is widely used in many fields, such as psychology, sociology, market research, business, and government. SPSS, originally named the Statistical Package for the Social Sciences®, was first developed by graduate students in 1968 to address a growing need for statistical tools that could be performed by computers. SPSS started as a small university-owned tool and quickly became adopted more widely by businesses and government entities. It has since undergone major revisions in software, corporation, and ownership, and it has recently acquired smaller companies, incorporating their products into their line, such as SYSTAT (statistical software), Deltagraph (graphing), and Quantime (market research). SPSS produces versions of their software for Windows, Mac, and Linux operating systems, with frequent version updates and a separate student version as well.

## Capabilities

There are two primary ways to use SPSS, with the Graphic User Interface (GUI) dialog boxes or with SPSS syntax, scripts, and macros. The GUI offers ease of use and a shallow learning curve, whereas the latter options offer greater flexibility, automation, and access to features that are not otherwise available through the GUI. For users who wish to transition from the GUI to syntax, any action permitted by a dialog box can be easily translated into an SPSS syntax window by using the "paste" button in a dialog box. In addition to enhanced flexibility, SPSS syntax also allows users to create an archive of the analyses previously conducted and allows users to recalculate previous analyses without clicking through all the options in the dialog boxes.

The SPSS GUI has three fundamental windows: the data editor, the syntax window, and the output window. The data editor uses a spreadsheet ("data view") and integrated variable descriptor editor ("variable view") for data management. The syntax window allows users to save and execute commands generated by the dialog boxes or to create specialized code for organizing and analyzing data. The results of analyses are organized and shown in the output window. All dialog box operations are available from each of the three types of windows.

The built-in SPSS tutorials are useful reference tools that can be accessed through the help menu in any of the GUI windows. SPSS also includes a syntax reference that explains the syntax structure for most or all of the statistical operations that can be performed.

Included with SPSS are several sample data sets, such as portions of the 1991 U.S. General Social Survey data, which are included to help new users acclimate to the software by offering examples of how users can structure data sets for particular

analyses as well as for conducting practice analyses with appropriate data.

### *Data Management*

SPSS can accommodate large data sets that have many thousands of variables and cases. Traditionally, SPSS data files are arranged such that individual variables are in columns and cases (participants or observations) are in rows. In addition to the data view, where actual data are entered and stored, there is also a variable view, which functions as a code book. Each variable name can be expanded into a more descriptive name by using the label option in the variable view (earlier versions of SPSS placed an 8-character limit on variable names; current versions of the program limit variable names to 64 characters). Each variable must be assigned a format such as numeric, string, or date; for numeric variables, the type of measurement might also be specified (i.e., scale, nominal, and ordinal). SPSS also provides the opportunity to assign meaningful labels to numeric values of variables (e.g., 1 = male and 0 = female). Other options in the variable view spreadsheet mode include defining the number of decimals displayed, determining the width of the variable columns, and specifying missing variables. SPSS automatically recognizes empty cells as missing; however, sometimes it is useful to assign a unique code for missing data to make it easier to check for data entry errors or to suit the particular needs of a project. Data might also be entered and saved in a SPSS syntax file using the *data list* command.

Data can be entered into the spreadsheet manually or imported from other database or statistical programs, such as Microsoft Access, SQL Server, SAS, Microsoft Excel, and so on. Similarly, data might be saved in other formats, such as previous versions of SPSS, SAS, Excel, or Stata files, ASCII text, and comma- or tab-delimited files. This feature provides additional flexibility when working with SPSS, because data can easily be imported from or saved in an alternate format and used by another program. Also, since Windows version 14.0 was introduced, multiple data sets can be open at the same time, which allows users to work with multiple data sets with ease. Once data sets have been saved as SPSS data files, researchers can merge separate data sets into a single file for storage or analysis.

#### SPSS Data Manipulation

Data can be sorted by a variable or a by combination of variables for organization and ease of inspection. Particular analyses require different data representations in the spreadsheet. Multidimensional scaling (MDS), for example, typically requires a square matrix of similarity measures, whereas a regression analysis requires data in columns. SPSS's online help is useful for determining the specific data representation required for a statistical analysis. If necessary, cases (rows) can be converted into variables (columns) and variables can be converted into cases using the *transpose* function. This conversion can be useful for analyzing a data set for different types of analyses. For example, in a multivariate data representation, repeated measures would be represented as separate columns, whereas in a univariate representation, the same data are organized in a single column indexed by a second column using *dummy codes*. Additionally, statistical tests can be conducted on a defined subset of the data using the *select* (*filter*) command (e.g., only women) or conducted separately for subgroups using the *split file* command (e.g., separate tests for men and women).

Researchers can create new variables based on their existing data using the *compute* or *recode* commands. Using the *compute* function, it is possible to create new variables from existing variables using different mathematical commands (e.g., summing, finding the mean, and transforming data using a log function). *Recode* lets you specify a new *effect* coding scheme by overwriting a variable or creating a new variable; for example, a continuous variable can be recoded into discrete categories.

Missing data might be imputed (estimated) in a couple of different ways. SPSS can assign the overall mean, median, or mode for a variable to any missing cases, or it might replace missing data based on algorithms that estimate the missing value based on data for other cases in the data set. Imputation is useful because it retains more cases for consideration in analyses, whereas missing data are usually basis for removal from analyses.

### *Data Analysis*

The software incorporates many analytic methods, many of which are available only as add-on modules that are priced separately from the SPSS base package. Basic descriptive statistics are supported by the base package, such as measures of central tendency and variance. Central tendency and variance can be calculated or visualized using graphics such as box plots, stem and leaf, and so on.

As a first step, researchers can check the distributions of their data to ensure that they are meeting the assumptions of their planned analyses (e.g., normality, distribution of errors, heteroscedasticity, and outliers). Reliabilities and other scale statistics can be computed as a preliminary step to ensure that researchers have psychometrically sound measurement scales. Data visualization is often considered important at the initial stage.

After meeting assumptions of tests and ensuring scales are reliable, researchers have myriad options available to them for statistical analysis. SPSS supports nonparametric tests such as chi-square, Wilcoxon's, and binomial tests. More advanced tests such as resampling or permutation can be conducted using SPSS syntax.

SPSS can also calculate a variety of parametric analyses. Simple analyses, such as the *t* test, one-way analysis of variance (ANOVA), correlation, and linear regression are included with the base system, whereas other analyses can be only used after separate modules have been purchased and installed. SPSS can analyze data using many common statistical techniques, such a multiple regression, logistic regression, ANOVAs, analyses of covariance (ANCOVA), multivariate analyses of variance (MANOVAs), and factor analysis. Additional modules for purchase include regression, generalized linear models (including hierarchical linear model [HLM] and generalized linear model [GLM]), neural networks, and others.

When conducting statistical analyses, SPSS provides the value of the test statistic, the *p* value, and it can also produce effect sizes or estimate residuals. Analyses typically use two-tailed tests of statistical significance. For each parametric or nonparametric test, users also have the option to tailor each analysis to meet their needs, for example by specifying linear contrasts within factorial ANOVAs to compare specific groups or trimming multiple regression models to improve model fit. Furthermore, for most statistical tests, assumption tests can also be performed along with the analysis.

SPSS supports most data visualization methods that are typically used in the social sciences, such as histograms, scatterplots, and bar or line graphs. Users have some control over the presentation of the graphs; many aspects of the figures can be altered in the GUI interface or altered using SPSS syntax. In addition to specifying visualizations via commands, many graphs can also be edited within the output window. Visualizations might be exported as image files or copied to the clipboard and pasted into many word-processing or presentation software applications.

Analysis results are placed into the SPSS output window, and the analyses are organized chronologically in the left sidebar, though they might be reorganized manually. The left sidebar not only organizes overall tests, but also portions of output within each test, like assumption checks, estimated values, and values for the statistical test. Results can be pasted to other programs such as a word processor, saved as different file formats, and can probably be passed to other programs.

### *Automation and Integration*

Most commands performed by SPSS can be automated, so that little user input might be needed to perform a complicated series of commands once the programming is complete. Several automation methods are available. SPSS syntax is the most conceptually simple and easily learned. In addition to the commands provided by the SPSS software, many user-created macros exist to conduct specialty analyses. Recently, Python integration has been implemented, which allows another automation option for SPSS users. Python is a programming language distributed under an open-source software license.

## Limitations

There are several considerations of which researchers should be aware before adding SPSS to their research toolbox. Current versions (17, as of late 2009) require at least Microsoft Windows XP,

Apple OS 10.4, and Linux kernel 2.6.9.42. Apple computers are currently supported, but under versions 8–12 they were not. Support for Mac OS was reinstated in 2004. If using a PowerPC-based Mac, then SPSS version 13.0 or earlier must be used; if using an Intel-based Mac, then SPSS version 15.0 for Mac or higher must be used.

SPSS is certainly no replacement for sound statistical knowledge. Like many other computer-based statistical applications, it is easy to conduct analyses that a user might not understand, and it is also easy to perform many "exploratory analyses" in search of significant effects that might not have originally been hypotheses of interest. Researchers must be aware that this latter issue presents a problem of inflated alpha or Type I error, where they might discover significant effects because of chance alone and that are not replicable. Similarly, SPSS will run analyses even if statistical assumptions are not met. It is the user's responsibility to check that the statistical assumptions are met to ensure the validity of the analyses conducted.

The price of SPSS can be a barrier for some users, especially for those interested in more complex analyses that require the purchase of additional modules. Several open-source alternatives such as R and PSPP do not have all the features and usability of SPSS but are viable and in development.

*P. Matthew Bronstad and Amanda R. Hemmesch*

***See also*** R; SAS; Statistica; SYSTAT

**Further Readings**

Einspruch, E. L. (2004). *Next steps with SPSS*. Thousand Oaks, CA: Sage.

Levesque, R. (2008). *Programming and data management for SPSS statistics 17.0: A guide for SPSS statistics and SAS© users* (5th ed.). Chicago: SPSS, Inc. Retrieved February 2009, from http://www.spss.com/statistics/base/ProgDataMgmt_SPSS17.pdf

## Standard Deviation

In the late 1860s, Sir Francis Galton formulated the law of deviation from an average, which has become one of the most useful statistical measures, known as the standard deviation, or *SD* as most often abbreviated. The standard deviation statistic is one way to describe the results of a set of measurements and, at a glance, it can provide a comprehensive understanding of the characteristics of the data set. Examples of some of the more familiar and easily calculated descriptors of a sample are the range, the median, and the mean of a set of data. The range provides the extent of the variation of the data, providing the highest and lowest scores but revealing nothing about the pattern of the data. And, either or both of the highest and lowest scores might be quite different from the scores in between. The median is the single middle score (or average of the two middle scores) that tends to be used to compensate for unusually high or low scores, but again, it still provides little information about the overall characteristics of the data. The mean is probably the most familiar, as well as easily calculated, descriptor of a given set of numbers. As individuals, we like to know where we fit in comparison to most other people. Countless surveys have been conducted, for example, to determine what "most" people earn or the taxes that most people pay. It is also useful to gauge a school's educational success by determining whether test scores are above or below average. On a more personal level, it can be interesting to know the average height of the players on one basketball team compared with another. However, although the mean is quick and easy to calculate, it is also subject to distortion by even a single extreme score, or outlier, in the data set. For example, the mean of a set of data can be the same; the mean height of two teams of basketball players might both be 7 feet tall. But, the variability of the players' heights on each team could be quite different. The height of players on team 1 ranges from 6 feet 9 inches to 7 feet 3 inches. The height of team 2 might realistically cluster around 6 feet tall, but one of their players is 7 feet 6 inches tall. A comparison of the means alone hides the fact that most of team 2 is much shorter than team 1. To obtain insight as to which team might be more likely to win tonight's game, more information is needed about the true variability of the players' heights.

To understand that variability, statisticians calculate the standard deviation, which is especially

**Table 1** Group Example of Data Distribution, the Mean, and the Difference of Standard Deviations

| | *IQ Scores* | |
|---|---|---|
| *Participant* | *Group 1* | *Group 2* |
| 1 | 95 | 95 |
| 2 | 95 | 97 |
| 3 | 97 | 99 |
| 4 | 97 | 99 |
| 5 | 100 | 100 |
| 6 | 100 | 100 |
| 7 | 100 | 100 |
| 8 | 105 | 103 |
| 9 | 105 | 103 |
| 10 | 106 | 104 |
| Mean | 100 | 100 |
| Standard deviation | 4.1 | 2.8 |

important to research because, although the other measures described previously are useful, the standard deviation provides a more accurate picture of the distribution of measurements. This statistic is an indicator of the distance of individual measurements from the mean score. A low standard deviation indicates that the data points are clustered tightly around the mean value, whereas a high standard deviation indicates that the data are less precise and spread across a large range of values.

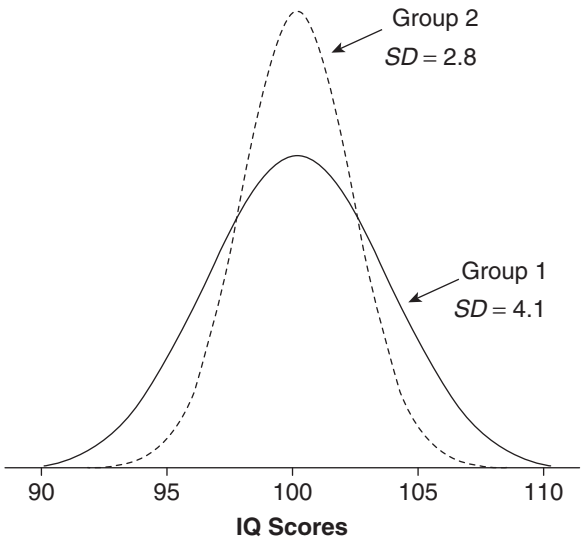



**Figure 1** Differences in the Normal Curve as a Result of Differences in Distribution of the Data Are Illustrated for the Example of Groups 1 and 2

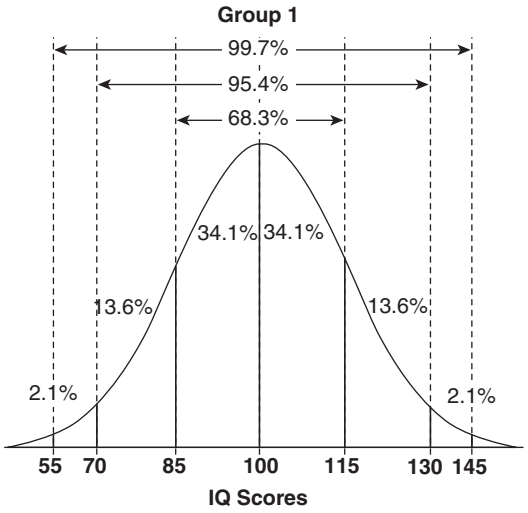



**Figure 2** The Relationship of the Standard Deviation and the Normal Distribution of Data as Illustrated by the Group 1 Data Set

For example, Table 1 provides the IQ scores of two samples. In these two samples, both groups have the same mean score of 100; however, the variability of group 2's data around that mean is about 30% less than the variability of group 1. In this case, a single standard deviation for groups 1 and 2 equals 4.1 and 2.8 IQ points, respectively. This means 2 standard deviations are twice that amount (i.e., 8.2 and 5.6), and 3 standard deviations are three times that amount (i.e., 12.3 and 8.4). Applying these data to a normal curve illustrates the difference in the dispersion of the two samples (see Figure 1). The data from group 1 cluster more tightly around the mean, producing a tall narrow curve, whereas the greater dispersion of data from group 2 produces a shorter broader curve by comparison.

Unlike the closely related measure variance, the standard deviation is expressed in the same units as the data. In the example, the data are IQ scores from two samples; therefore, the standard deviation is a number that represents IQ points. As shown in Figure 2, the standard deviation is specifically divided into three segments: three standard deviations both above and below the mean value. In a normal bell-shaped curve, each standard

$$\sqrt{\frac{\sum(X-M)^2}{N-1}}$$

Where $X$ = The individual score
$M$ = The mean of the sample (the group of scores)
$N$ = The sample size

**Figure 3** Steps, Combined Into a Single Equation, That Are Needed to Calculate the Standard Deviation

deviation represents a percentage of the sample scores, which is occasionally referred to as the $68-95-99.7$ rule. That is, when normally distributed, 34.1% of the scores will fall above the mean score and 34.1% will fall below the mean for a total of 68.3% (accommodating for rounding error) of the scores within $\pm 1$ standard deviation from the mean. The second standard deviation encompasses 13.6% of the scores such that 95.4% are within $\pm 2$ standard deviations from the mean. Last, the third standard deviation encompasses 2.1% of the sample scores, making up a total of 99.7% within $\pm 3$ standard deviations from the mean. This leaves only 0.3% of scores falling at the two extremes of the normal curve, which means only 3 scores of 1,000 will range beyond $\pm 3$ standard deviations from the mean. So, roughly two thirds of the scores in a normal distribution are within one standard deviation of the mean.

In mathematical terms, Figure 3 represents a compilation of steps, combined into a single equation, that are needed to calculate the standard deviation. However, this equation can be broken down into seven steps (but only six simple calculations). Knowing the individual steps for calculating the standard deviation can be useful for removing the mystery and fear surrounding this number, and for understanding its purpose. These steps are as follows:

1. Calculate the mean of your data set
2. Subtract that mean from each of the scores in your data set to determine the individual deviation of each score from the mean
3. Square each of those individual deviations
4. Sum all of the squared deviations
5. Divide that sum by one less than the sample size $(N-1)$
6. Last, take the square root of that result
7. That final number equals the standard deviation of your data set

Because most data sets are much larger than the example here of two groups of 10, it is fortunate that most spreadsheet and statistical programs include a function that will easily calculate the standard deviation of a set of data. The computer calculation will provide this information for between-group comparisons or for repeated measurements within the same group(s). These calculations are also typically provided as an item contained in the results of other more in-depth statistical analyses (e.g., $t$ tests or analysis of variance).

Although the description and ease of calculating the standard deviation makes this single number seem like the answer to understanding all the important aspects of the data, there are also key limitations to interpreting precisely what this number means. One of the most important considerations is to know that the standard deviation provides information only about the measurements taken from the sample group(s) that was selected as representing all the possible measurements that could ever be made of that same type. The earlier example of IQ scores is a case in point. The two samples selected in Table 1 are not truly representative of IQ scores of an entire population. The actual standard deviation has been determined by sampling very large numbers of people, testing large numbers of different samples, and testing samples representing a wide range of ages and other characteristics that would be typical of the population as a whole. Therefore, the standard deviations in Table 1 are only an estimate of the true variability within that population. The accuracy of that estimate increases depending on the sample selection and how well the sample represents the population of interest. The accuracy is also highly subject to the size of the sample studied. Larger samples tend to have less variability and be somewhat more representative of an entire

population. Somewhat related to the importance of the size of the sample is the number of different samples that are tested. Testing many smaller samples, fewer larger samples, or best of all, testing many large samples, will all improve the accuracy of the estimate of the true variability in a population. The important thing is, as long as these limitations are understood, the standard deviation is still a valuable statistic important to understanding characteristics of the data and to interpreting the results of testing.

*Dawn M. Richard*

***See also*** Degrees of Freedom; Mean; Median; Normal Distribution; Population; Range; Sample Size; Sampling; SAS; SPSS; Standard Error of the Mean

**Further Readings**

Galton, F. (1869). *Hereditary genius: An inquiry into its laws and consequences*. London: Macmillan. Retrieved February 23, 2009, from http://galton.org/books/hereditary-genius/galton-1869-Hereditary_Genius.pdf

Hampton, R. E. (1993). *Introductory biological statistics* (2nd ed.). Dubuque, IA: William C. Brown.

Lane, D. M. (2009). *Standard deviation and variance*. HyperStat online statistics textbook. Retrieved February 25, 2009, from http://davidmlane.com/hyperstat/A16252.html

Motulsky, H. (1995). *Intuitive biostatistics*. New York: Oxford University Press.

Norman, G. R., & Streiner D. L. (2000). *Biostatistics: The bare essentials* (2nd ed.). Hamilton, Ontario, Canada: BC Decker.

# Standard Error of Estimate

Standard error is the estimated standard deviation of an estimate. It measures the uncertainty associated with the estimate. Compared with the standard deviations of the underlying distribution, which are usually unknown, standard errors can be calculated from observed data.

First, this entry will explain how estimates are obtained. Then, the entry discusses how the standard errors of estimates are derived, with an emphasis on the differences between standard errors and standard deviations. Two examples are used to illustrate the calculation of standard errors of a parameter estimate and standard errors of a future outcome estimate, respectively. Finally, the relationship between standard errors and confidence intervals is discussed.

## Estimate

One main purpose of research designs is to make inferences about a population. Many population characteristics are described by numerical measures or can be translated into numerical values. Numeric descriptive measures of a population are called parameters of the population. Common examples of parameters include the percentage of voters supporting candidate Mr. Jones, the average weight of 1-month-old babies, the variance or standard deviation of the annual income of fresh college graduates in year 2008, the correlation coefficient between annual income and years of education, and the increase of average annual income associated with 1 more year of education. In many cases, it is infeasible to sample the whole population, and the sampled information might not be 100% accurate; thus, the true values of the parameters are unknown and need to be estimated. An estimate is an approximation of the true parameter value using specific rules when sampled information is incomplete, noisy, or uncertain. The rule to calculate an estimate is called an estimator, which could be a formula or a procedure. Different estimators give different estimate values for the same parameter, and many statistical methods have been developed to evaluate the good and bad qualities of estimators.

Another main purpose of research design is to make predictions for future events. Often, future events can be characterized by numeric values, such as the highest temperature of an area next week, the life expectancy of someone who has already lived to 80, and the cure rate if a new treatment is applied. The true values of the measurements of future events are unknown until the event happens. Statistical regression models study the relationship between the specific measurements (outcome) and other related variables (predictors) based on historic data, and then estimate the future values. The estimate of a future value is called a prediction. Although the true values of the

regression coefficients are fixed, different models give different predictions for the same future value. Even with the same model the predictions will vary because the regression coefficients, which measure the relationship between the predictors and the outcome, are estimated from a particular sample and vary from sample to sample.

## Standard Error of Estimate

The differences between the true value and its estimates are called errors. The exact value of an estimate depends on the estimating procedure as well as the sample used to calculate the estimate. Because the sampled data are a random sample of the underlying population, the estimates vary from sample to sample even when the same estimating procedure is used, especially when the sample size is small. When the sample size goes to infinity, the estimate converges to its expectation, which might or might not be the same as the true value being estimated. The difference between the expectation of an estimate and the true value is called bias. Some estimating procedures produce unbiased estimates, whereas others do not. Both randomness in the sample and bias lead to difference between estimates and the corresponding true value, that is, errors.

The standard deviation of the estimate for a parameter or a future value can be illustrated with the following steps. First, sample a population multiple times. Second, calculate the estimate based on each sample. Third, take the difference between the estimates and the true value, thus a sample of errors is obtained. When the number of samples goes to infinity, the distribution of the errors can be observed. The standard deviation of the errors is the positive square root of the variance of the errors, that is, the expectation of the squared difference between the error and the mean of the error distribution.

In real life, sampling infinite times is infeasible. The error distribution is usually not observed, and the standard deviation of the errors is unknown. Thus, many statistical methods have been developed to estimate the standard deviation of the errors. The estimated standard deviation of the errors in an estimating procedure is called the standard error of the estimate, which is sometimes abbreviated as standard error or $SE$ or $S_E$. It needs to be pointed out that standard errors are only estimates of the standard deviation of errors, not the standard deviation itself. For example, when an estimate is divided by its standard error instead of its standard deviation, the resulting test statistic follows a Student $t$ distribution instead of a normal distribution, which will be explained in more detail using the examples in the following two sections. When sample size is large enough, the Student $t$ distribution converges to the normal distribution and standard errors converge to standard deviations as a measure of the uncertainty. Here, whether the sample size is "large enough" or not depends on the particular estimating procedure and the parameter or future measurement being estimated.

Given the standard error of an estimate, then the standard error of its linear functions can be calculated accordingly.

$$SE(X + c) = SE(X)$$
$$SE(cX) = cSE(X),$$

where $X$ represents the estimate and $c$ is a constant.

## Standard Error of Estimates of Parameters

Take the standard error of the mean estimate as an example. First, notations are set up. A sample of $n$ independent observations, $X_1, X_2, \ldots, X_n$, is taken from a population with mean $\mu$ and variance $\sigma^2$. A widely used estimate of the population mean is the average of all observations

$$\overline{X} = \frac{1}{n}\sum\nolimits_1^n X_i,$$

where $i = 1, \ldots, n$. The mean estimate $\overline{X}$ is normally distributed with mean $\mu$ and variance $\sigma^2/n$ if the distribution of $X_i$ is normal. However, the assumption that $X_i$ follows normal distribution is not always necessary in deriving the distribution of $\overline{X}$. When the sample size is moderately large ($n > 30$), by central limit theorem, $\overline{X}$ from independent identically distributed samples converges to a normal distribution with the same mean and variance $\sigma^2/n$. Through some transformation, $(\overline{X} - \mu)/(\sigma^2/\sqrt{n})$ converges to a standard normal distribution (mean zero and variance one).

The standard deviation of the population is $\sigma$, which can be estimated by the sample standard deviation, which is denoted as $s$ in the following equations. That is, the standard deviation based on the observed sample, which serves as an estimate of the population standard deviation $\sigma$, is

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}.$$

The standard error of the mean estimate can be calculated as

$$SE(\overline{X}) = \frac{s}{\sqrt{n}}.$$

Note that the standard error of the mean estimate is different from the standard deviation of the mean estimate,

$$SD(\overline{X}) = \frac{\sigma}{\sqrt{n}}.$$

Here, $SD(\overline{X})$ will not be known unless $\sigma$ is known. When people substitute $SD(\overline{X})$ with $SE(\overline{X})$ in calculating$(\overline{X} - \mu)/SD(\overline{X})$, the resulting statistic $(\overline{X} - \mu)/SD(\overline{X}) = (\overline{X} - \mu)/(s/\sqrt{n})$ follows a Student's $t$ distribution with mean zero, variance one, and degree of freedom $n - 1$, instead of a standard normal distribution. Analysts cannot replace the standard deviation with the standard error directly because additional uncertainty is added when an estimate of the standard deviation is used instead of the true standard deviation itself.

## Standard Error of Future Outcome Estimates

Take the simple linear regression with one predictor as an example. Suppose the expectation of $Y$, $E(Y)$, and $X$ have a linear relationship,

$$E(Y) = \beta_0 + \beta_1 X,$$

where $\beta_0$ and $\beta_1$ are regression coefficients with fixed but unknown true values. When a series of $X_i, Y_i$, $i = 1, \ldots, n$, is observed, because of the randomness associated with each individual $Y_i$, the linear relationship becomes

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

where the residuals $\varepsilon_i$ are assumed to follow a normal distribution with mean zero and constant variance $\sigma^2$. The least squares estimators of the regression coefficient, $\hat{\beta}_0$ and $\hat{\beta}_1$, are obtained by minimizing the sum of the squares of residuals. The explicit forms for the least square estimators are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}.$$

Note that although $\beta_0$ and $\beta_1$ are constants, $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimated from the sample and vary from sample to sample, that is, $\hat{\beta}_0$ and $\hat{\beta}_1$ are random.

For future observations with $X = X^*$, the corresponding $Y$ values can be estimated as $\hat{Y}^* = \hat{\beta}_0 + \hat{\beta}_1 X^*$. The error of the prediction $\hat{Y}^*$ is calculated as $\hat{Y}^* - Y^*$, where $Y^* = \beta_0 + \beta_1 X^*$ is the expected value of the future outcome corresponding to $X^*$. Through some algebra, it can be proved that the standard deviation of the errors is

$$SD(\hat{Y}^*) = \sigma\sqrt{\frac{1}{n} + \frac{(X^* - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}}.$$

Because $\sigma^2$ can be estimated by the sample mean square error (MSE)

$$S^2 = \sum_{i=1}^{n}\{Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)\}^2/(n-2),$$

the standard error of $\hat{Y}^*$can be calculated as

$$SE(\hat{Y}^*) = S\sqrt{\frac{\frac{1}{n} + (X^* - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}}.$$

As pointed out earlier, $(\hat{Y}^* - Y^*)/SD(\hat{Y}^* - Y^*)$ follows a standard normal distribution, whereas $(\hat{Y}^* - Y^*)/SE(\hat{Y}^* - Y^*)$ follows a Student's $t$ distribution with degrees of freedom $n - 2$.

### Express Standard Errors as Confidence Intervals

Standard errors provide measurements of uncertainty in an estimate and are widely used. In the previous examples, the estimate divided by its standard error follows a Student's $t$ distribution, and the uncertainty of the estimate can also be expressed with confidence intervals. For example, when $(\overline{X} - \mu)/SE(\overline{X})$ follows a Student's $t$ distribution with degrees of freedom $n - 1$, the 95% confidence interval for $\mu$ is $(\overline{X} - SE(\overline{X})t_{.975, df=n-1}, \overline{X} + SE(\overline{X})t_{.975, df=n-1})$, where $t_{.975, df=n-1}$, is the 97.5th quantile of the Student's $t$ distribution with degree of freedom $n - 1$. Because $\overline{X}$ and $SE(\overline{X})$ are random and depend on data, the 95% confidence interval is also random and vary from sample to sample. If sampling is repeated multiple times and the 95% confidence interval is constructed each time, 95% of the intervals will cover the true $\mu$ value.

If the data are sampled from a normal distribution or the sample size is moderately large ($n > 30$), the confidence intervals based on standard errors can use quantiles from normal distributions. For example, for large $n$, $(\overline{X} - \mu)/SE(\overline{X})$ converges to a standard normal distribution. Then, the 95% confidence interval for $\mu$ is calculated as $(\overline{X} - SE(\overline{X})Z_{97.5}, \overline{X} + SE(\overline{X})Z_{97.5})$, where $Z_{97.5}$ is the 97.5th quantile of the normal distribution. Specifically, interval $(\overline{X} - SE(\overline{X}), \overline{X} + SE(\overline{X}))$ covers true $\mu$ 68% of the time, interval $(\overline{X} - 1.96SE(\overline{X}), \overline{X} + 1.96SE(\overline{X}))$ covers true $\mu$95% of the time, and interval $(\overline{X} - 3SE(\overline{X}), \overline{X} + 3SE(\overline{X}))$ covers true $\mu$ almost all the time.

When the distribution of the estimate divided by its standard error is unknown, conservative confidence intervals for the true value can be calculated, where the standard errors are derived using relationships such as Chebyshev's inequality.

*Qing Pan*

***See also*** Bivariate Regression; Central Limit Theorem; Confidence Intervals; Error; Estimation; Mean; Parameters; Sample; Standard Deviation

#### Further Readings

Wackerly, D. D., Mendenhall, W., III, & Scheaffer, R. L. (2003). *Mathematical statistics with applications* (6th ed.). Pacific Grove, CA: Duxbury.

## Standard Error of Measurement

The standard error of measurement is a statistic that indicates the variability of the errors of measurement by estimating the average number of points by which observed scores are away from true scores. To understand standard error of measurement, an introduction of basic concepts in reliability theory is necessary. Therefore, this entry first examines true scores and error of measurement and variance components. Next, this entry discusses the standard error of measurement and its uses. Last, this entry comments on methods for reducing measurement error.

### True Score and Error of Measurement

When psychological tests are administered, points on items measuring the same construct are aggregated to generate the raw score for each respondent. The raw score, or observed score, can be conceptualized as the sum of two distinct components, the true score and the error of measurement. The true score is caused by systematic influence that produces consistency to the observed score on the test. Note that the true score does not speak to the true standing of an individual on the construct; instead, it only indicates the unchanging portion of the observed score. The error of measurement is a result of chance influence that produces random variation to the observed score on the test.

$$\text{Observed score} = \text{true score} + \text{error of measurement}$$

The concepts of the true score and the error of measurement can be illustrated with the next example. Consider an examinee who takes an ability test and assume that the same test can be administered multiple times without him or her remembering the test items. The examinee's observed score might be influenced by factors such as his or her true ability level, test-taking skills, lucky guessing, and fatigue. The portion of the observed score contributed by the ability level and the test-taking skills is invariant across different possible administrations of the test, as long as his

or her ability and test-taking skills remain unchanged. This consistent portion of the observed score is defined as the examinee's true score. Note that the true score might be influenced by his or her level of the construct, that is, his or her ability level, as well as factors that are irrelevant to his or her level of the construct, such as his or her test-taking skills.

In contrast, fatigue and lucky guess might vary on different administrations of the test, so their effects are transient rather than systematic. The portion of the observed score associated with these transient factors is the error of measurement. The error of measurement is random, being positive sometimes and negative other times, and when the same test is taken by the examinee for a large number of times, the average of the error will approach zero. In other words, the average of those observed scores will approximate the true score.

## Variance Components

The amount of variation in observed scores on a test from a population of interest can also be partitioned into two components: variation caused by true scores and variation caused by an error of measurement. When expressed in terms of variance components:

$$\text{Observed score variance} = \text{true score variance} + \text{error variance}$$

The reliability of a test is defined as the extent to which the observed score variance is caused by true score variance. With the same amount of observed score variance, the larger the true score variance, or the smaller the error variance, the more reliable a test is. Reliability can be expressed as follows:

$$\begin{aligned}\text{Reliability} &= \frac{\text{true score variance}}{\text{observed variance}} \\ &= 1 - \left(\frac{\text{error variance}}{\text{observed variance}}\right).\end{aligned}$$

Reliability can be understood as the degree to which a test is consistent, repeatable, and dependable. The reliability coefficient ranges from 0 to 1: When a test is perfectly reliable, all observed score variance is caused by true score variance, whereas when a test is completely unreliable, all observed score variance is a result of error. Although the reliability coefficient provides important information about the amount of error in a test measured in a group or population, it does not inform on the error present in an individual test score. The standard error of measurement is used to determine the effect of measurement error on individual test scores so as to aid decision making on individual outcomes. Serving in a complementary role to the reliability coefficient, it also functions as another important indicator of the psychometric quality of a test or measure.

## Standard Error of Measurement

The standard error of measurement is the standard deviation of error of measurement in a test. It is closely associated with the error variance, which indicates the amount of variability in a test administered to a group that is caused by measurement error. The standard error of measurement (*SEM*) is expressed as

$$\begin{aligned}SEM &= \sqrt{\text{error variance}} \\ &= \sqrt{\text{observed variance} \times (1 - \text{reliability})} \\ &= \text{standard deviation of observed scores} \\ &\quad \times \sqrt{1 - \text{reliability}}.\end{aligned}$$

The standard error of measurement is a function of both the standard deviation of observed scores and the reliability of the test. When the test is perfectly reliable, the standard error of measurement equals 0. When the test is completely unreliable, the standard error of measurement is at its maximum, equal to the standard deviation of the observed scores. An additional advantage of the standard error of measurement is that it is in the original unit of measurement. With the exception of extreme distributions, the standard error of measurement is viewed as a fixed characteristic of a particular test or measure. Whereas the Pearson product-moment coefficient measure of reliability is commonly used for the calculation of the standard error of measurement, others have argued that the intraclass correlation coefficient is a more

appropriate one to use. This is additionally complicated by the fact that different formulas can be used to calculate the intraclass correlation coefficient. Still others have pointed out that the standard error of measurement can be calculated from the square root of the mean square error term in a repeated-measures analysis of variance (ANOVA).

Given that the overall variance of measurement errors is a weighted average of the values that hold at different levels of the true scores, the variance found at a particular level is called the conditional error variance. The square root of the conditional error variance is the conditional standard error of measurement, which can be estimated with different procedures.

In the previous example of the test taker, with multiple administrations of the same test without the memory and practice effect, his or her observed test scores will form a normal distribution with the mean at the true score, and the standard deviation of his or her observed test score will approximate the standard error of measurement. This is an approximation because the standard error of measurement is estimated on a group instead of on an individual. Indeed, the estimate is an overall value for the whole test, and researchers have shown that standard error of measurement can be different at different levels on the scale score (i.e., conditional standard error).

### *Probabilistic Statements Around True Scores*

With the standard error of measurement, probabilistic estimates can be generated around true scores, using the normal curve and the standard normal table. Given a true score, confidence intervals can be constructed around the true score to estimate the probable range of observed scores. When a given true score is determined as the cutoff score for decision-making, individuals with true scores at or above the true score level can nevertheless obtain observed scores below this cutoff. To account for the effect of measurement error, confidence intervals can be created to estimate the cutoff for observed score so that one can ensure that a certain percentage of individuals (e.g., 68%, 95%, etc.) whose true scores are higher than the true score cutoff will be selected. The confidence interval can be created using the following formula:

$$\text{Confidence interval} = \text{true score} \pm z \times SEM,$$

where $z$ is the $z$ score associated with the desired probability level, which is available in standard normal tables. For example, when an IQ true score is determined as the cutoff of eligibility for an academic program, and the standard error of measurement is known in the relevant population, setting the cutoff observed score to one standard error below the cutoff true score can ensure that 68% of test takers whose true scores are above 100 will be selected.

Using standard error of measurement, one can also estimate the likelihood with which an observed score comes from a given true score by calculating the corresponding $z$ score, using the following expression:

$$z = \frac{(\text{observed score} - \text{true score})}{SEM}.$$

After obtaining the $z$ score, one can look up the probability associated with the $z$ score from a standard normal table. For example, when an observed score is two standard errors higher than a true score of interest, the $z$ score associated with the probability is 2, and there is only around a 2.5% chance that the observed score came from an individual having the said true score.

### *Probabilistic Statements Around Observed Scores*

When an observed score is known, confidence intervals might be placed around the observed score to estimate the probability of obtaining a certain true score. For example, one might wish to find out the probable range of a test-taker's true score given his observed score. The correct estimate of the confidence interval around observed scores, presented by Harold Gulliksen, is slightly more complex and does not involve standard error of measurement. However, the standard error of measurement is often used to place confidence intervals around observed scores in practice, despite cautionary notes by some authors against such use. Gulliksen suggested the use of Observed score $\pm$ $z \times SEM$ to estimate true score, noting that this interval provides "reasonable limits" of true score estimates if precise probability statements are not made. The use of standard error of measurement to create intervals around observed scores can be

a good approximation if the reliability is reasonably high and if the observed score is close to the mean of the reference group. Given these two conditions for approximation, it is somewhat misleading when confidence intervals are constructed around the observed score to create probabilistic estimates of the true score.

### *Difference Between Observed Scores*

When testing the significance between two observed scores, the standard error of measurement is not applicable because difference scores are less reliable than single scores. Instead, the standard error of measurement of score difference is used:

$$SEM_{\text{difference}} = \text{Standard deviation of observed scores}\sqrt{2 - \text{reliability}_1 - \text{reliability}_2}.$$

If the difference between two observed scores is more than two standard errors for score difference, one can be 95% confident that the true scores are different between these two test takers.

## Typical Use of Standard Error of Measurement

The standard error of measurement is used in industrial and organizational psychology in three typical ways. An applicant's test score can be compared with a cutoff true score to determine the likelihood of obtaining a score above or below the cutoff after retest. Scores between two applicants can be compared with each other to determine whether they are different from each other. Given the means of two groups, the difference between the means can also be compared using the standard error of measurement.

In the educational arena when high-stakes testing (e.g., college admissions or determination of qualification for special education services) is involved and test scores are used for classification purposes, errors of measurements are highly consequential. In such circumstances, it would be useful to report conditional standard error of measurement near the cutoff scores or critical values used for such decisions.

The Standards for Educational and Psychological Testing has recommended that both overall and conditional standard error of measurement be reported when test information is disseminated. In addition, these standard errors of measurements need to be provided in raw scores and original scale units as well as each derived score to be used in interpretation.

## Preventing Measurement Error

Jum Nunnally and Ira Bernstein have indicated that it is usually better to prevent measurement errors from occurring in the first place than trying to assess their magnitude and effects after the fact. They suggest some useful steps to reduce measurement error, which include writing items that are clear, ensuring that test instructions are easily understood, and adhering to the prescribed conditions for test administration. Other recommended steps include making subjective scoring rules as explicit as possible and investing in the training of raters to do their jobs well.

*Frederick T. L. Leong and Jason L. Huang*

***See also*** Analysis of Variance (ANOVA); Confidence Intervals; Intraclass Correlation; Pearson Product-Moment Correlation Coefficient; Reliability; Standard Deviation; True Score; Variance; *z* Score

### Further Readings

American Educational Research Association. (1999). *Standards for educational and psychological testing.* Washington, DC: Author.

Dudek, F. J. (1979). The continuing misinterpretation of the standard error of measurement. *Psychological Bulletin, 86,* 335–337.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

Harvill, L. M. (1991). Standard error of measurement. *Educational Measurement: Issues and Practice, 10,* 33–41.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Urbina, S. (2004). *Essentials of psychological testing.* New York: Wiley.

# Standard Error of the Mean

The standard error of the mean is a number that represents the variability within a sampling

$$\frac{SD}{\sqrt{n}}$$

Where $SD$ = The standard deviation of the sample
and $n$ = The size of the sample

**Figure 1** Formula for Calculating Standard Error of the Mean

distribution. In other words, it is how measurements collected from different samples are distributed around the mean of the total population from which the sample measurements could be taken. The mean of each sample of measurements, therefore, is an estimate of the population mean. Although the means from each sample provide an estimate of the mean of the entire population, measurements collected from numerous samples are quite unlikely to be equivalent. This raises the question of how well each sample mean represents the true population mean. That accuracy can be estimated by calculating the standard error of the mean, which provides an indication of how close the sample mean is likely to be to the population mean.

In random samples of the population, the standard error of the mean is the standard deviation of those *sample* means from the population mean. So strictly speaking, the functional difference between the two measures is that the standard deviation is used to calculate the variability of *individual measurements* around the mean of a sample, and the standard error of the mean is used to calculate an estimate of the variability of *the means of multiple samples* around the mean of the population. The formula for calculating the standard error of the mean (see Figure 1) is closely related to that of the standard deviation (for calculation, see Figure 3 in Standard Deviation entry) in that it is the standard deviation divided by the square root of the sample size. And, like the standard deviation (see Figure 2 in Standard Deviation entry), the means of 68.3%, or about two thirds, of all the sample means will be within $\pm 1$ *standard errors* of the mean (rather than $\pm 1$ standard deviation), 95.4% will be within $\pm 2$ standard errors of the population mean, and 99.7%, or almost all the sample means will be within $\pm 3$ standard errors from the population mean (only 0.15% will be more than three standard errors either above or below the population mean).

For example, assuming the height of all individuals in a total population of 330 people was measured, the mean of those measurements would be

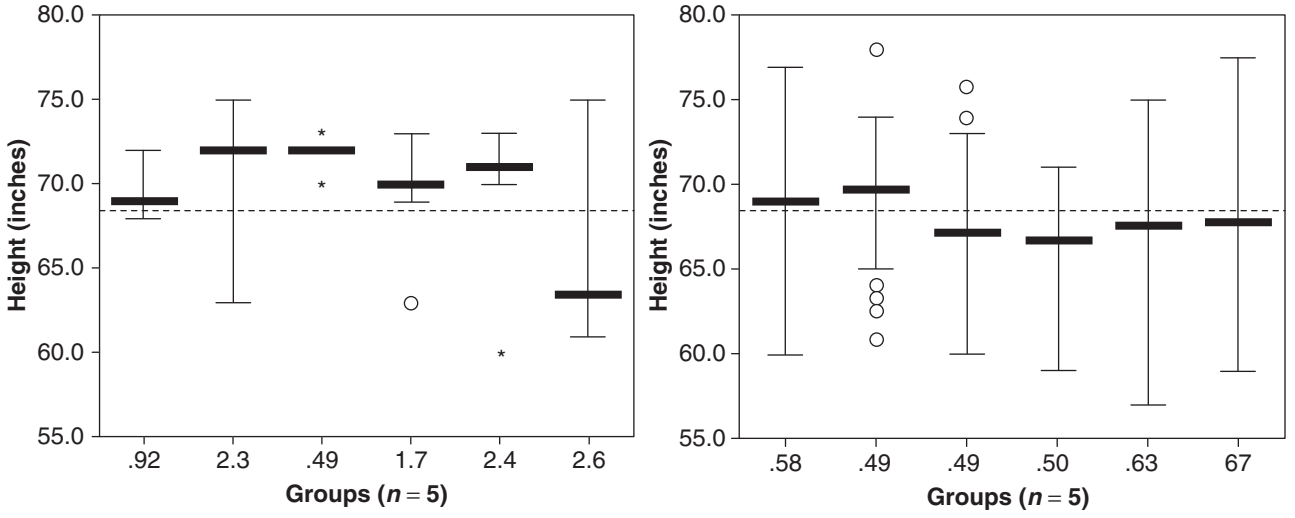



**Figure 2** The Effect of Sample Size on Accuracy of Estimating the Population Mean

*Note:* The population mean = 67.7 inches tall (dashed line).

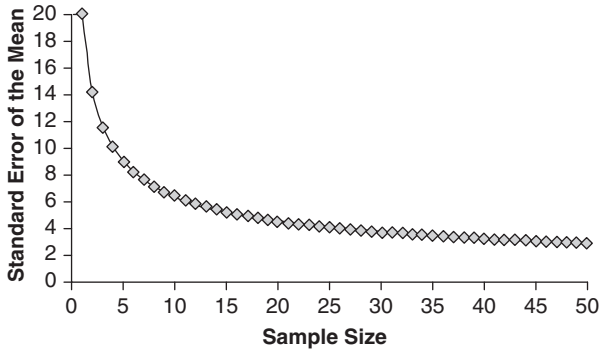



**Figure 3** The Function of Sample Size on the Standard Error of the Mean

*Note:* Holding the SD constant at 20.

the true mean of that entire population. Because it is rarely feasible to conduct an experiment that tests an entire population (and entire populations are typically much larger than this fictitious example), experimenters collect data from smaller samples that are randomly selected from the population. The size of the sample will influence the accuracy of the estimate of the true population mean. Looking at Figure 2, the graph on the left shows data for 6 samples of 5 individuals each and the graph on the right shows data for 6 samples of 50 individuals each. For both graphs, each vertical line (with the cross bars on the ends) shows the spread of the data in that sample (asterisks and circles represent extreme values within each sample), and the heavy horizontal line is the mean height of the individuals in each sample. The horizontal dotted line on each graph illustrates the location of the true mean of the total population of 330 individuals. The means of each of the small samples (left side, $n = 5$) deviate from the true population mean much more than the means of each of the larger samples (right, $n = 50$). The means of the smaller sample sizes also show much more variability in the distance from the population mean. On both graphs, the calculated estimates of the deviation of those sample means from the population mean (i.e., the standard error of the mean) are shown just above the $x$-axis. Comparing the standard error calculations on each graph, it is easy to observe that the larger samples provide a more consistent and truer estimate of the population mean than the smaller samples. The larger samples deviate from about one half to two thirds of a standard error from the mean, whereas the smaller sample sizes vary from one half to more than two and a half standard errors from the mean. Clearly, when feasible, testing a larger sample will provide a better estimate of the population, although there are practical limits.

Another way to view the effect of the sample size on the standard error of the mean is to calculate the standard error holding the standard deviation constant (numerator, Figure 1) while increasing the size of the sample (denominator, Figure 1). For example, if the standard deviation is held constant at 20, and the standard error of the mean is calculated for sample sizes from 1 to 50 (see Figure 3), when the sample size is small there is a large difference in the estimate of the standard error of the mean. Small samples, as noted previously, do not necessarily provide a good estimate of the population mean and increasing the sample size can have a considerable effect on increasing the accuracy of the sample mean as an estimate of the population mean. However, as the sample size continues to increase, the difference in the standard error begins to level off. Increasing the sample size no longer has much effect on how well the sample represents the population and the benefits of a larger testing sample are diminished.

It bears repeating that the standard deviation is the deviation of individual observations from the mean of the sample, and the standard error is the deviation of sample means from the mean of the population. The standard error of the mean and the standard deviation are closely related and many consider the two as equivalent terms; as long as one is reported, plus the sample size, the other can be easily calculated. Although this is true, others argue against using the standard error of the mean to represent variability when publishing data because it is the standard deviation that describes the actual variability within the sample being reported, and the standard deviation is, by definition (see Figure 1), larger than the standard error of the mean. This argument contends that it is misleading to report the smaller standard error of the mean, which is only an estimate of how well the sample represents the true population mean. Whereas it is essential not to lose sight of the fact that the initial sample can only be an estimate of

the population characteristics, as long as the distinctions between the two statistics are understood, both can provide important information about the characteristics of the sample and the population from which it was drawn.

*Dawn M. Richard*

***See also*** Normal Distribution; Population; Sample Size; Sampling; Standard Deviation

**Further Readings**

Lane, D. M. (2009). *Sampling distribution: Simulations and demonstrations (Interactive)*. HyperStat Online Statistics Textbook. Retrieved March 30, 2009, from http://onlinestatbook.com/stat_sim/sampling_dist/index.html

Lane, D. M. (2009). *Standard error of the mean.* HyperStat Online Statistics Textbook. Retrieved March 30, 2009, from http://davidmlane.com/hyperstat/A16252.html

McDonald, J. H. (2008). Standard error of the mean. In *Handbook of biological statistics* (pp. 99–103). Baltimore, MD: Sparky House Publishing.

Moore, D. S., & McCabe, G. P. (2005). *Introduction to the practice of statistics* (5th ed.). New York: W. H. Freeman.

Nagele, P. (2003). Misuse of standard error of the mean (SEM) when reporting variability of a sample. A critical evaluation of four anaesthesia journals. *British Journal of Anaesthesia, 90,* 514–516.

# STANDARDIZATION

*Standardization* is a term that is used in variety of ways in research design. The three most common uses of the term relate to standardization of procedure, standardization of interpretation, and standardization of scores.

## Standardization of Procedure

Standardization of procedure is highly important in any kind of research design and essentially refers to experimental control. This is relevant for any test or experimental procedure and includes the standardization of instructions, administration (including manipulation), and measurement of variables of theoretical interest. Instructions need to be as clear as possible for the particular population. The presentation order of test material or experimental manipulations needs to be identical between people, trials, or conditions. Time constraints might be specified. In many test designs, instructions might also include ways of handling questions of participants. For some tests, oral instructions might include the consideration of rate of speaking, tone of voice, inflections, facial and bodily expressions, or pauses. For example, correct answers might be given away if smiling or pausing when a critical test item is being read out loud. In essence, this type of standardization tries to reduce the influence of any extraneous variable on the test or experimental performance of participants. A lack of standardization of procedure will lead to unreliable results, which in turn will threaten the validity of the test or experiment. The standardization of procedure as a term is more often used in the context of test development (e.g., test instructions, item order, and time limits) than in experimental design. In experimental designs, researchers are concerned with internal and external validity, which is often treated as a separate topic.

## Standardization of Interpretation

The standardization of interpretation refers to the standardized interpretation of obtained scores. This is of particular importance in test administration and interpretation. Here, the concept of so-called norm groups is central. Psychological tests have no predetermined standards against which performance of individuals or groups of individuals can be evaluated. Therefore, scores are typically compared with some norm, which is the performance of others obtained using the same test. Norms imply some form of normal or average performance. For example, if the typical 8-year-old child answers 10 questions of 50 correctly, then this would be the norm for 8-year-olds. Without knowing the average performance of a comparable group, the raw score of 10 correct answers is meaningless. The raw scores on tests could include the number of correct answers (or percentage of correct answers), number of errors or mistakes, reaction times, mean or modal response to some attitude item, or some other objectively measurable indicator relevant for the content of the test. In addition to such age norms, percentile ranks or

grade norms are other typical examples. To obtain these norms, tests are administered to large and representative samples drawn from the population for which the test is being designed. This is an essential element of test standardization. The quality of the test norm depends on this so-called standardization sample chosen for norming the test. For example, if a test is developed to measure mental abilities for managers, it would be meaningless to norm the test on patients who underwent psychiatric treatment.

The utility and applicability of norms is hotly debated. Norms are important in many applied contexts (e.g., in mental health, educational, and work settings), where decisions need to be made about individuals falling above or below some particular normative criterion (e.g., for selecting individuals for therapy, students for university entry, or new employees likely to meet performance targets). However, norms often change over time and can be highly context specific. The Flynn effect (named after James Flynn) is the most well-documented case for the temporal instability of norms. IQ scores have been steadily increasing over the decades, requiring constant renorming of IQ tests. Using outdated norms could lead to invalid inferences about individuals. Similar problems exist in regard to the applicability of norms across culturally and ethnically diverse populations. The interpretation of test scores can be highly misleading if inappropriate norms are used. In many societies, minority groups do not have the same access to education and professional development. If norms developed for majority group individuals were applied to minority group individuals, then inappropriate conclusions would be drawn if the test performance is influenced by these differences in educational opportunities (or other socioeconomic variables that put minority group members at disadvantage). Some critiques have therefore called for local norming (establishing norms based on specific local populations), and some have even suggested an abandoning of the use of test norms altogether. Although some of the criticism of norms is justified, abandoning norms would be counterproductive and problematic in practical settings. It would also open the door for more subjective and potentially discriminatory practices, which are the very issues that critics of norms are concerned with.

## Standardization of Scores

The third use of standardization is in relation to individual or group scores. This is somewhat associated with the previous usage of the term, because one particular way of norming scores is to express raw scores in the form of $z$ scores (sometimes called standardized scores). Individual scores are expressed in units that indicate the position of the individual relative to the distribution of scores in his or her group. A score of 0 would indicate that the individual has a score that is exactly at the mean of the group, whereas a score of 1.0 would mean that the individual scores one standard deviation above the mean for this group. Like scores based on some population norms, these $z$ scores allow normative interpretations. Nevertheless, the ad hoc sampling of participants (typically using relatively small samples of university undergraduate students) together with often non-normal distributions of psychological variables in small samples precludes meaningful comparisons of $z$ scores across studies and samples. One alternative is to compute so-called percent of maximum possible (POMP) scores, which express raw scores in terms of the maximum possible score. Any score can be converted into POMP scores by taking the raw score minus the minimum score and then dividing it by the possible scoring range. If test scores had binary response options, this would be equivalent to the percentage of correct answers. If multiplied by 100, the converted scores range between 0 and 100. This scoring method standardizes the scores so that they can easily be compared across alternative scoring methods, populations, and instruments. They have the advantage of conveying immediate meaning. Many psychological tests have different answer scales and it is often not clear how a score of 3 on a scale from 1 to 4 compares with a score of 3 on a scale from 0 to 4. This arbitrariness of scaling hinders efforts to build a more cumulative science. However, POMP scores do not address the relative meaning of scores compared to some real or statistical norm. For example, a score of 55 only indicates that the individual could have reached a score of 100 but only achieved 55. It does not indicate whether 55 is a particularly high or low score on that construct.

### *Standardization to Adjust for Response Styles*

Another application of standardization of scores is common in the psychometric and more recently in the cross-cultural literature, in particular if subjective evaluations of attitudes, personality, values, beliefs, or some other psychological construct are being measured using some form of rating scale. It deals with the same problem of score comparability, but this time the concern is the presence of response styles or biases that might distort observed scores. The two most commonly discussed response styles are acquiescent responding (also called yay-saying and its opposite nay-saying) and extreme responding (also called modesty responding). Acquiescent responding results in some upshifting or downshifting of mean scores independent of item content, whereas extreme responding is the selection of more extreme scores (either acceptance or rejection) than would be expected based on the true level of the underlying construct. Such tendencies can be expressions of individual differences or can be characteristics of cultural groups, indicating some cultural response or communication style. If individuals or groups are more likely to agree with items or endorse more extreme statements than they actually hold, this will certainly create problems in the interpretation of scores. In practical contexts, this might lead to the invalid selection of candidates in educational or occupational settings or incorrect diagnoses on self-rated mental health checklists. In research contexts, it might obscure real between-group differences, or more typically, will result in differences that are spurious and not related to the construct of interest. For example, it has been observed that Japanese participants typically use the middle category of a response scale, which is independent of item content. If we were to compare Japanese and U.S. samples (that typically do not show the same response style), many significant differences might be found reflecting this difference in response styles, but not differences in the measured psychological construct. It can also lead to paradoxical findings if the scores of different subgroups with different response styles are combined and the data are being analyzed in the whole sample without considering the different score means in the subgroups. Imagine a researcher has measured the frequency of laughing each day and the number of times individuals cry during the week. There is a negative relationship in each group separately (the more you cry, the less you laugh). However, one group gives higher responses irrespective of item content (assuming that the overall emotional responsiveness is not different between groups). If the data from the two groups are pooled, the relationship between crying and laughing might become zero or even positive because of the different positions of the two groups. Therefore, ignoring response styles might lead to paradoxical relationships between constructs. Several standardization or score adjustment procedures have been discussed in the literature.

The first procedure has already been discussed, namely $z$ transformation. If scores are $z$ transformed in each group, it effectively removes any mean differences between the two groups and might circumvent the problems described in the crying/laughing example. This will work well in situations where researchers are conducting structure-oriented tests, that is, tests examining relationships between constructs using correlational techniques (correlation, regression, factor analysis, path modeling, etc.). However, this will not work if researchers are interested in examining means or true levels of the underlying variable (so-called level-oriented tests, e.g., $t$ test and ANOVA). In this case, using $z$ transformation will remove any between-group differences (because the mean will be standardized to zero in each group separately). It will also not work if there are individual differences in responding.

A second strategy that has been used frequently, especially in the personality and cross-cultural literature, is within-subject standardization or ipsatization. Each individual's score on each variable will be adjusted for the total mean across all variables for that person. This adjusts for overall personal tendencies to respond more positively or negatively, adjusting for acquiescence responding of individuals. If the standard deviation for each individual is used for adjustment, then it might also control for extreme responding. Therefore, each score is then expressed as the relative standing of that person on that variable in relation to all other variables that are used for within-subject standardization. The overall score for that person across all items is consequently zero. Using the

crying versus laughing example again, if we standardized the responses of individuals, the adjusted score after ipsatization could be interpreted as the relative frequency of laughing compared with crying (and vice versa). This has important theoretical and statistical implications. Theoretically, using within-subject standardization or ipsatization, a researcher would assume a limited resource, or fixed-size pie scenario. This might or might not make sense in particular areas of investigation. Using the crying/laughing example, we would assume that all people have the same level of emotionality and spend all their time either crying or laughing. In motivational research, it might be logical to assume that individuals have only a limited area of energy, and therefore there is a balance or trade-off for where this energy can be invested. However, a different motivation researcher might assume that some individuals have higher levels of motivation overall and would like to predict what personality variables are associated with this overall level of motivation. In this case, ipsatized scores would not be useful.

Another problem is that if all the scores are influenced by an unmeasured third variable, it might also obscure any differences and lead to paradoxical results after standardization. Standardizing the length of the body features of a mouse and an elephant might lead to the conclusion that the tail of the mouse is longer than that of the elephant. However, because both are related to the overall size of the animal, this statement is certainly incorrect (if scores are interpreted as absolute measures). This form of standardization therefore raises important theoretical questions about the meaning and interpretation of such scores.

Within-subject standardization is also problematic mathematically because it creates ipsative scores. It results in mean scores that are constant for each individual, and each score for an individual is linearly dependent on scores on all other variables, while being independent of the scores of other individuals. This implies mathematically that the sum of variances and covariances is zero in each row and column of the covariance matrix. This will create problems for factor analysis and other multivariate tests that require creation of an inverse matrix. This standardization also forces at least one of $k - 1$ (with $k$ being the number of variables) covariance terms in each row and column to be negative, independent of the substantive relationship. This can have important theoretical consequences, because at least one negative correlation in a correlation matrix is not determined by the relationship between constructs or scores but by pure methodological artifacts. Furthermore, both the average correlation between all variables (which will be negative) and the relationship between standardized scores and external criterion variables (which will always sum up to zero) are known. The implications for standard statistical tests are wide-ranging. Reliabilities are affected, factor analyses might yield uninterpretable loading patterns, and mean scores can be seriously overestimated or underestimated. Some of these effects are mitigated if a large number of unrelated constructs is used in research. Some guidelines say that ipsatization might work if it is based on at least ten unrelated constructs. However, from a practical perspective in that case, there would be little need for within-subject standardization. Using nonparametric tests, many of these aforementioned problems can be overcome. However, nonparametric tests are not as widely used, are less well understood, and are often less powerful and stringent. There have also been a lot of recent efforts to overcome some of these problems with more advanced multivariate methods (such as structural equation modeling). However, the results are somewhat mixed and it needs to be determined to what extent ipsative scores can be used with modified multivariate tests.

Another type of score adjustment that has been used primarily in the cross-cultural literature is within-culture standardization. Here, the score across all items and individuals within a group is used for adjusting each score of each individual within a sample. The idea is to remove cultural response styles. Again, if the standard deviation is used in addition to (or instead of) means for adjustment, the adjusted scores might also control for cultural extreme responding. This method raises somewhat similar theoretical problems, if the constructs are highly related. However, the statistical implications are less severe, especially if larger numbers of variables are being used.

Sometimes, double-standardized scores have been advocated in the cross-cultural literature. Here, the scores are first ipsatized and then these

ipsatized scores are corrected using within-culture standardization. This strategy will remove both personal and cultural response styles, but the meaning of the resulting scores can be ambiguous. It also creates problems associated with ipsatized scores, and multivariate statistical tests might not run properly.

A final but less frequently recommended and used method is to use covariate analyses to adjust for the overall response tendency of the individual. Similar to within-subject standardization, the overall mean across all items or scores for an individual is created and then partialled in an analysis of covariance or partial correlation. This use of the covariate analysis has many of the same problems as ipsatized scores. In addition, it might create problems of singularity (where the criterion variable is perfectly related to the predictor variable, in this case the covariate variable is partialled first).

The use of these techniques has increased during the last three decades, even when accounting for the increased publication rate. It certainly reflects the increased trend to use self-reports in psychological and social research, and it also demonstrates an increasing awareness among researchers about the problems of using raw scores to make cross-cultural or cross-ethnic comparisons. Nevertheless, the use of these score adjustment procedures is not without problems, and researchers should exercise caution when using them. For research purposes, it is better to model such response style factors and estimate the effect directly rather than controlling for it without knowing the exact extent and nature of it. It is also possible to use within-subject designs to control for individual differences in reaction to stimuli.

*Ronald Fischer*

***See also*** External Validity; Internal Validity; Likert Scaling; Psychometrics; Rating; Raw Scores; Response Bias; Sampling; Standardized Score; Threats to Validity; Validity of Measurement; Within-Subjects Design; *z* Score

### Further Readings

Cohen, P., Cohen, J., Aiken, L. S., & West, S. G. (1999). The problems of units and the circumstance for POMP. *Multivariate Behavioral Research, 34,* 315–346.

Fischer, R. (2004). Standardization to account for cross-cultural response bias: A classification of score adjustment procedures and review of research in JCCP. *Journal of Cross-Cultural Psychology, 35,* 263–282.

Hicks, L. E. (1970). Some properties of ipsative, normative and forced-choice normative measures. *Psychological Bulletin, 74,* 167–184.

Murphy, K. R., & Davidshofer, C. O. (1994). *Psychological testing: Principles and applications.* Englewood Cliffs, NJ: Prentice Hall.

## Standardized Score

There are times when it is important to compare the scores of different types of data that are scored in different units or to compare scores within a sample or a population. One common example is that students typically want to understand how their test score compares with the scores of the rest of the class. Or they might want to understand how their scores compare across different classes, which could each be scored using somewhat different methods. In either case, the challenge is somewhat like attempting to compare apples with oranges. A standardized score is calculated on an arbitrary (but universal) scale, which has the effect of turning the apples and oranges into pears, scores that can be more easily evaluated and interpreted. In other words, a raw score can be converted from one measurement scale to another to facilitate data comparison. This entry discusses comparative and standardization methods.

### Comparative Methods

Several different methods are used to create scores that are comparable with one another; however, some of the more familiar methods can have limited comparative accuracy. For example, there are 23 students who are each taking mathematics, history, and psychology midterm exams. One student, Ben, scores a 53 in math, 77 in history, and 88 in psychology. Comparing these scores, it seemed to Ben that he is not very good at math. But these are raw scores that can tell us only the number of questions that were correctly answered on each examination. Without a frame of reference, it is not possible to determine Ben's standing in each

class relative to the other students or to understand the relationship of his performance across his different classes. One way to understand Ben's performance compared with the other students is to use the range of examination scores within each class. In this example, the scores of all the students who took the math examination ranged from a low of 48 to a high of 57. This means Ben's score was just above the middle of that range, which might indicate his score is close to average when compared with the rest of the class. The history exam scores ranged from 75 to 99, so Ben's score was nearly at the bottom of that range, seeming to be a very poor score compared with the other students. And, the psychology scores ranged from 86 to 90, which puts Ben's score precisely in the middle of the range of his classmates' scores.

A somewhat more useful way of calculating scores for comparative purposes would be to calculate the percentage scores for each examination (i.e., number of correct answers divided by total possible correct answers, then multiplied by 100). This type of calculation might provide more equitable scores for a comparison within and between classes. Because the percentage scores (not to be confused with percentile scores) are dependent on the total number of questions on the examination, converting Ben's raw score of 53 on his math examination to a percentage might show that his score compares differently both within and between his classes than when using simple range values for comparison. Some might immediately think that calculating percentages solves the comparison problem: not necessarily and, most likely, not adequately for accurately understanding Ben's performance within each class, or his overall performance when comparing across his classes. Ben's math exam had a total of 150 questions, so his raw score of 53 converts to 35%. His history test had 100 questions, so his raw score of 77 remains at 77%. And, his psychology test had 90 questions, turning his raw score of 88 to 98%. When comparing scores calculated as a percentage of the total instead of comparing raw scores, it seems that Ben's math score is even worse than he originally thought. *But,* Ben's math instructor creates very difficult exams and on this exam everyone in the class scored from 32% to 38%. Consistent with the comparison of the range of scores, his score of 35% falls precisely in the center of the class percentages, which could mean he is actually an average student. On the history exam, the other students scored from 75% to 99%. Ben's score of 77% is well below average, and it is still consistent with his ranking using the range of scores. But, on the psychology exam the other students scored from 92% to 100%, which indicates Ben's score of 98% is close to the top of the class, a fairly substantial increase from the comparison using the range of scores. However, that there are differences between the two comparison methods makes it very difficult to determine which set of comparisons are correct. Furthermore, while there are indicators of the spread of the scores from each exam using the upper and lower ends of either the raw or the percentage scores, neither method provides information about the variability of the scores. If only one student scored lower than Ben on the math exam and the rest scored very close to the top of the range, Ben's score would not be average after all. Although his score was *numerically* in the middle of both the range and percentage scores, he actually got the second lowest score of all the 23 students. And, because the range and the percentage scores are determined only within each group, both types of scores are insufficient for comparing Ben's scores from each of the classes. These simple examples illustrate some of the important pitfalls of using some of the possible methods of comparisons.

## Standardization Methods

One way to avoid the problems described previously is to use standardized scores when comparing both within and across data sets. The standardized (or standard) score is a dimensionless quantity that can be calculated from any set of scores to transform them into equal units that facilitate a comparison within a particular group of scores or across different types of scores. The transformation process is called standardization, and it is one method for normalization of a data set. Although this calculation can be used for a small sample as in these midterm examination examples, the larger the sample size, the more accurate the standardized score and the most accurate score is calculated from a population mean and standard deviation. However, in this case, Table 1 provides all the necessary information for calculating the standardized score based on the sample means of each of Ben's classes. A standardized score is easily calculated by

**Table 1** Examples of Midterm Examination Scores and Different Scoring Methods Used for Comparative Purposes

| | *Midterm Examinations* | | |
|---|---|---|---|
| | *Math* | *History* | *Psychology* |
| Number of students taking examination | $n = 23$ | $n = 23$ | $n = 23$ |
| Number of examination questions | 150 | 100 | 90 |
| Range of all students' raw scores | 48–57 | 75–99 | 86–90 |
| Ben's raw score (# of correct answers) | 53 | 77 | 88 |
| Ben's percentage score | 35% | 77% | 98% |
| Mean raw score of the class | 51 | 85 | 88 |
| Standard deviation of scores from the mean | 2.10 | 6.45 | 1.48 |
| Ben's standardized $z$ score | 0.95 | −1.24 | 0.00 |
| Ben's standardized scaled score | 12.85 | −13.72 | 10 |
| Ben's standardized $T$ score | 59.5 | 37.6 | 50.0 |

subtracting the mean of the raw scores of the data set from an individual raw score. That difference score is then divided by the standard deviation of the raw scores within that data set (see Figure 1a). Applying this formula to Ben's scores (using the mean and standard deviation from Table 1) provides a more accurate picture of his overall standing among his fellow students (see Figure 1b). The result of this standardization calculation is often called the $z$ score (or normal score). The $z$ score has a mean of zero and a standard deviation of 1. Applying the $z$ score calculation to all of Ben's examination scores (see Figure 2) shows that Ben is almost one standard deviation above the mean (1 unit above average) on his math midterm examination, which indicates he performed much better than the other comparative methods (previously) suggested. Ben's psychology midterm score is exactly on the mean, or average, and his history midterm score is more than one standard deviation below the mean (negative $z$ score). But, if Ben goes home to tell his parents about his midterm exams, it is unlikely that his parents will understand if he tells them he got a zero in psychology. That is where alternative standardization scores are useful.

$$z = \frac{X - M}{SD}$$

Where $z =$ The standardized $z$ score

$X =$ The individual raw score

$M =$ The mean of the sample

$SD =$ The standard deviation

For example,

Ben's Math Exam $z$ score is :

$$.95 = \frac{53 - 51}{2.1}$$

**Figure 1** Formula for Calculation of the Standardized z Score

Several alternative standardized scores are provided here, each of which is simply an extension of the $z$ score calculation (see Figure 2). The scaled score is calculated by applying a simple linear transformation to the calculated $z$ score. The scaled score equals $3(z) + 10$, a mean of 10 and standard deviation of 3. The $T$ score (not be confused with the $t$ test that has an entirely different purpose) was arbitrarily decided to have a mean of 50 and a standard deviation of 10; therefore, the formula to transform the $z$ score is $T = 10(z) + 50$. Using a $T$ score, it is a little easier for Ben to explain his average midterm score to his parents as a score of 50. Along with the $z$ scores, Table 1 provides the standardized scaled scores and $T$ scores for each of Ben's midterm examinations, and Figure 2 illustrates how these two scores are related to the $z$ score. And then there is probably the most well-known example of this type of transformation, the standardized scores for IQ tests. Reporting a $z$ score for the IQ score would be very upsetting to most parents because an average IQ

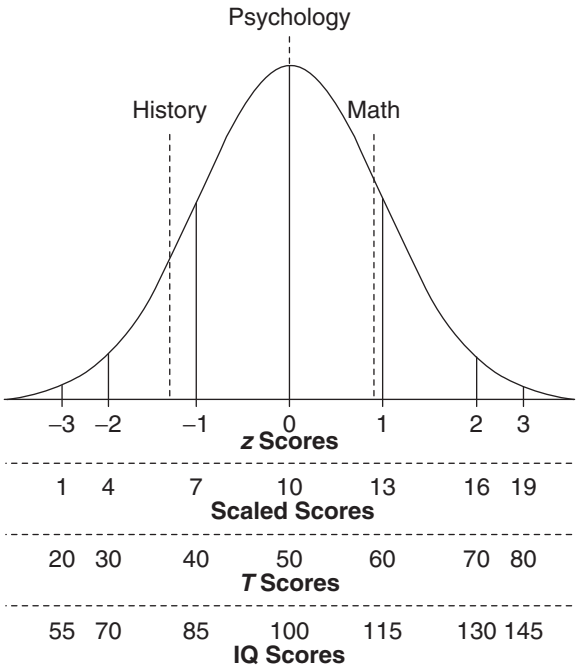



**Figure 2** The Normal Curve With Comparisons of Four Standardized Scoring Methods: *z* Scores, Scaled Scores, *T* Scores, and IQ Score

*Note:* The example midterm scores are indicated and apply to the *z* score and *T* score axis labels only.

would be zero, so applying a linear transformation where the mean is 100 and the standard deviation is 15 (see Figure 2) results in a more understandable score. The transformation of IQ scores can be calculated as IQ = 15(*z*) + 100 (Wechsler IQ), or IQ = 16(*z*) + 100 (Stanford-Binet IQ, not shown). Looking at Figure 2, the normal curve helps to understand where a particular score is in relationship to other scores, but there are also probability tables for each of the different standardized scores that can be used to find the corresponding percentile score. For example, looking up any of Ben's *z*, scaled, or *T* scores on the appropriate probability table (available on the Internet) can provide a percentile rank that is a more concrete way of understanding how many people scored above and below Ben's score. Because the scaled score and *T* score are linear transformations of the *z* score, the tables for all of these standardized scores provide the same percentile rank of 82.9 for Ben's math score, 10.6 for his history score, and 50.0 for his psychology score. Thinking back to the use of percentages, Ben's worst score, his math midterm examination, was only 35%, which was clearly not a good representation of Ben's mathematical ability. And, his best score, 98% in psychology, was really only average.

Scores resulting from different sets of data can take many different forms, and whereas there are some calculations that are commonly used to compare scores (e.g., percentage scores), they are not necessarily the most appropriate choice for accurate comparison of scores. The standardized score will provide the best picture of how scores compare both within and between groups of data. The standardizations discussed here, the *z* score and the related scaled score, *T* score, and IQ score, are some of the most well-known and widely used standardization methods that allow for the most accurate comparisons of scores, although each rests on the original calculation of the simple *z* score. By examining these different transformations of the standardized score, it is easy to understand that other alternative transformations could be created to provide a score on yet another scale. Although the example of midterm examinations used here represents a very small sample ($n = 23$) and rather skewed sets of scores, it is sufficient to illustrate the *concept* of standardized scores. It is important to remember that the larger the sample size, the more likely it is that the calculated *z* score will represent the most accurate probability of the population.

*Dawn M. Richard*

***See also*** Mean; Normal Distribution; Normalizing Data; Percentile Rank; Population; Probability, Laws of; Range; Sample; Standard Deviation; *z* Distribution; *z* Score

**Further Readings**

Association of Chief Psychologists with Ontario School Boards. (2009). *Conversion table for derived scores [standardized scores]*. Retrieved April 30, 2009, from http://www.acposb.on.ca/conversion.htm

Geographical Techniques Online. (2003). *Table of probabilities of the standard normal distribution [for z-scores]*. Retrieved April 30, 2009, from http://techniques.geog.ox.ac.uk/mod_2/tables/z-score.htm

Lyman, H. B. (1971). *Test scores and what they mean* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.
Moore, D. S., & McCabe, G. P. (2005). *Introduction to the practice of statistics* (5th ed.). New York: W. H. Freeman.

## STATISTIC

S*tatistic,* a singular noun, has three meanings: (1) an observed number or event, (2) a mathematical operation (and the resulting number) performed on sample data, and (3) an operation (and the resulting number) obtained from a sampling distribution. The plural form for each of these three cases is *statistics*. A fourth meaning of the word *statistics* is a branch of mathematics devoted to the analysis, interpretation, and presentation of empirical data. Statistics, as a branch of mathematics, is a singular noun. In most contexts *statistic* (with an s) refers to this branch of mathematics, which scientists and others use to analyze data from empirical studies.

In conversational use, *statistic* means a single number or a single event. The 84 in *84 people attended the reunion* is a statistic. *Statistic* sometimes conveys a sense of the impersonal. *His death was just a statistic* uses *statistic* in the same way as Josef Stalin's famous quote to Winston Churchill, "A single death is a tragedy; a million deaths is a statistic."

In the discipline of statistics, a descriptive statistic is a mathematical operation that captures some characteristic of a sample of observations. Each operation has a formula and, of course, each formula has a name. For a sample, the mean ($\Sigma X/N$) is a descriptive statistic, as is the median, standard deviation, and the 95th percentile, each of which has its own formula. Not only is the formula itself called a statistic, but the resulting numerical value can be referred to as a statistic. Thus, the expressions, $M = 98.2°$F, standard deviation ($SD$) = 12 points, $z = -1.50$, and 95% confidence interval (CI) = 0.50, −2.25 each qualify as a statistic and are often referred to as sample statistics. The corresponding population characteristics are called *parameters*. Although almost all research reports focus on statistics calculated from samples, the purpose of research is to discover relationships among populations. Thus, the most satisfactory statistics are those that provide mathematically unbiased estimates of parameters. The sample mean qualifies; the sample mode does not.

A *statistic* is also a value from a sampling distribution. A sampling distribution is a collection of sample statistics derived from a mathematical formula rather than empirical observations. Both the graphical form of the collection and the mathematical formula that produces the collection are called sampling distributions. Sampling distributions that researchers commonly use include inferential statistics such as the $t$ distribution, chi-square distribution, and the $F$ distribution. Thus, $t$, $\chi^2$, and $F$ values calculated from data are inferential statistics. In the same way that an expression such as $M = 98.2°$F or $SD = 12$ points is called a statistic, $t = 2.03$, $\chi^2 = 3.84$, and $F = 1.84$ are each called inferential statistics. The relationship of inferential statistics to research design is an intimate one. "An adequate experimental design involves not only a satisfactory plan for conducting the trial, but also an appropriate statistical method for evaluating the results. The two can ill be divorced" (Snedecor, 1946, pp. xv–xvi).

In general discourse, the word *statistic* is not a neutral word, and the connotation that accompanies it can be negative or positive, sometimes strongly so. Context and past usage determine the connotation. The phrase, *just a statistic,* conveys a negative tone of dismissal. The common expression, *you can prove anything with statistics,* is similarly dismissive. The phrase *lies, damned lies, and statistics,* attributed to Disraeli but found first in the writings of Mark Twain, even equates statistics with falsehood.

On the positive side, the phrase, *statistically proven,* carries a tone of assurance. A statistically proven conclusion is one that is based on evidence, according to the speaker. A prediction made with *statistical certainty* is one that can be relied on.

Using the word's connotation in an argument is a common rhetorical device. However, citing actual statistical results is usually more effective and often can prevail against an argument that rests only on connotation.

*Chris Spatz*

***See also*** Descriptive Statistics; Parameters; Sampling Distributions

**Further Readings**

Howell, D. C. (2008). *Fundamental statistics for the behavioral sciences* (6th ed.). Belmont, CA: Thomson Wadsworth.

Kirk, R. E. (2008). *Statistics: An introduction* (5th ed.). Belmont, CA: Thomson Wadsworth.

Snedecor, G. W. (1946). *Statistical methods applied to experiments in agriculture and biology* (4th ed.). Ames, IA: Collegiate Press.

Spatz, C. (2008). *Basic statistics: Tales of distributions* (9th ed.). Belmont, CA: Thomson Wadsworth.

Spatz, C. (2008). Statistical techniques and analysis. In S. F. Davis & W. Buskist (Eds.), *21st century psychology: A reference handbook* (pp. 46–54). Thousand Oaks, CA: Sage.

## STATISTICA

Statistica is a data mining and analysis program sold by StatSoft, Inc., of Tulsa, Oklahoma. The program is broad in scope with separate modules for advanced or specialty features oriented to data-intensive industries such as power production, pharmaceuticals, insurance, finance and banking, health care, and electronics. The program is available for individual computers, enterprise solutions for multiple users, and as a Web-based program. This entry describes features that are most apt to be of interest to readers of this encyclopedia, such as typical analyses, user interface, graphics, and documentation.

The basic module contains standard frequentist analysis techniques such as descriptive statistics; analysis of variance (ANOVA)/multivariate analysis of variance (MANOVA), including post hoc tests and planned comparisons; multiple regression; correlations; cross-tabulation analyses; basic nonparametric analyses; and distribution fitting statistics. An advanced module adds numerous additional analysis methods. A partial list of these methods includes generalized linear models, time-series forecasting, survival and failure time analysis, structural equation modeling, log linear analysis, and power analysis. There are also quality control statistics such as those associated with the Six Sigma movement and root cause analysis. It also contains multivariate exploratory methods such as principal components and factor analysis, cluster analysis, discriminant analysis, multidimensional scaling, and reliability/item analysis. There are other modules that permit neural network analyses and data mining.

Although Statistica is a highly sophisticated analysis system, it does not offer any current options for dealing with missing data other than casewise or listwise deletion and using means to impute missing values. There are some additional alternative missing data imputation methods in the time-series analysis routine.

For many psychological investigations, it is necessary to calculate effect sizes, both for interpretation and to allow later inclusion of the study's results in a meta-analysis. Effect sizes for mean differences obtained through ANOVA/MANOVA are partial eta-squared. These rely on the amount of variance in the dependent variable that was explained by the independent variable. There are several variations of effect sizes (the *g* and *d* family of effect sizes) that express the results as difference between means in standard deviation units. These are often used in meta-analyses. Statistica does not produce these effect sizes, although they can be relatively easily calculated from the statistics reported by Statistica.

Bayesian analyses are limited to certain data mining techniques and neuro network analyses, which are inherently Bayesian in nature. Other analyses do not permit the introduction of prior distributions or probabilities, at least not through the graphical user interface. Bayesian methods are becoming increasingly common in statistics, but other programs will need to be used to perform them.

Statistica contains an extensive array of graphic techniques. Because the program is used in many industries, it has the capability of creating real-time graphs. For typical social science research, the usual graph types (line, histogram, scatterplot) are available, plus graphs useful for descriptive purposes such as variations on pie charts and similar depictions. A potentially useful variation on the scatterplot that is not available in some other software is the bag plot, which is an extension of the Tukey box plot to two dimensions, illustrating central location, skewness, correlation, and allowing for quick identification of outliers.

The successful use of a statistical analysis system requires knowledge of statistical methods and

expertise in computer or Web usage. However, StatSoft, Inc., who make Statistica, provides a variety of resources for those novice computer users with only a basic understanding of statistics. On the company's website, and in the Statistica "Help" menu, users will find the link to the Electronic Statistics Textbook that offers training in the understanding and application of statistics. Virtual chapters contain diverse topics including, but not limited to, experimental design, basic statistics, multivariate techniques, and advanced linear/nonlinear models. StatSoft, Inc. also provides links to various white papers covering the company's industry solutions (e.g., pharmaceutical, insurance, and marketing), products, data mining, and other technical content (e.g., analysis of a split-plot design and random number generation).

To help those new to statistics and this particular software package, the Statistica help menu contains an electronic manual, a glossary, a statistical advisor, and links to animated help videos. The electronic manual provides numerous resources beyond software usage, including general statistics information to help the user understand concepts and analyses. Some example topics include elementary concepts (e.g., what is a variable and why the normal distribution is important) as well as content regarding standard analysis techniques (e.g., *t* tests, ANOVA, and regression). The Statistica glossary represents another all-inclusive resource at the user's fingertips that contains information on adjusted means to the Z distribution.

Additionally, Statistica incorporates a statistical advisor that serves several purposes. First, the advisor might be used to supply users' information regarding various statistical methods (e.g., overviews, assumptions, and graphs). The advisor might also be used to assist the user in identifying the most appropriate analysis methods. A third purpose is to aid the user in using the software to conduct various procedures. The advisor functions by asking the user to respond to questions in a stepwise fashion, such as "Do you want to ..., Describe, summarize, tabulate data; or Test hypotheses (predictions) about your data ..." If the user indicates that he or she wants to test hypotheses, for example, a new set of questions is posed, such as "You can test hypotheses regarding ..., Differences between variables; or Relationships between variables ..." This process continues until the user is presented with information regarding the pertinent analysis (e.g., independent sample *t* tests, Pearson's *r,* and multiple regression).

The help menu also contains links to Web-based animated help videos on a variety of topics, including overviews pertaining to user interface, graphs, output types; and a how-to section covering group analyses, using data filtering and cleaning tools, and using similar tools.

Statistica has an interactive graphical user interface and allows the user to customize menus and toolbars. The software affords the user complete control over the level of complexity of the program's functions. Data files are typically displayed in spreadsheets. The basic form of the spreadsheet is a two-dimensional table arranged as cases (rows) and variables (columns). On the simplest level, Statistica spreadsheets might contain both numeric and text data. However, the spreadsheets also offer the flexibility of containing several types of media (i.e., audio, video, graphs, photographs, and pages of text). Furthermore, each cell (in the data file) can be extensively edited and formatted with common editing functions found in word-processing software. The spreadsheets might also be formatted to allow other users to access data files or reports. All the formatting and other editing options are found in the many toolbars presented in the graphic user interface (e.g., file, edit, view, insert, and format).

More advanced computer users and novices alike might choose to use Statistica's Visual Basic macros. The software provides three options for creating macros automatically, including an "analysis" and "keyboard" macro that will record the settings and output options chosen for a particular analysis. Setting up macros is necessary for real-time data collection, analysis, and display.

To conduct a statistical analysis, users are given numerous options found in the following menus: statistics, data mining, graphs, tools, and data. Each menu provides the user with a variety of functions. The standard analysis options (e.g., basic statistics, ANOVA, and regression) are found in the statistics menu, and each procedure is initiated with a tabbed dialog box. Statistica provides novice users with the "quick" tab, which includes only the basic options needed to perform a particular analysis. Advanced users who prefer a more detailed analysis would select the "advanced" tab

for complete options. Once an analysis is performed, the user is only a few mouse clicks away from creating detailed graphs (as described earlier).

Statistica provides three basic options to display results, specifically workbooks, reports, and standalone windows. They can be used in many combinations and can be customized in a variety of ways. Workbooks and reports can hold Statistica documents as well as other types of documents, including Microsoft Excel spreadsheets and Microsoft Word documents. Workbooks contain a two-panel view, including an outline that allows the user to navigate easily from one part of the analysis to another, as well as a document viewer that allows the user to view different aspects of the analysis side by side. Reports present the typical method of handling output where each object (e.g., a spreadsheet, graph, or other object) is displayed sequentially in a word-processor style document. Furthermore, all output objects (spreadsheets and graphs) can contain other embedded and linked objects and documents, so the output can be organized hierarchically in a variety of ways. Files might be saved individually (spreadsheet, workbook, report, etc.); however, the user might also save all documents as a new project file that maintains the current work session and might be opened at a later time.

*Jay C. Thomas and Paul G. Michael*

***See also*** Analysis of Variance (ANOVA); Bayes's Theorem; Data Mining; General Linear Model; Multiple Regression; Power; Principal Components Analysis; SAS; Software, Free; Structural Equation Modeling

**Further Readings**

Anderson, N. H. (2001). *Empirical direction in design and analysis*. Mahwah, NJ: Lawrence Erlbaum.

Grimm, L. G., & Yarnold, P. R. (Eds.). (1995). *Reading and understanding multivariate statistics*. Washington, DC: American Psychological Association.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7,* 147–177.

StatSoft, Inc. (2007). *Electronic statistics textbook*. Tulsa, OK: StatSoft. Retrieved May 30, 2008, from http://www.statsoft.com/textbook/stathome.html

Thomas, J. C. (2008). Statistical considerations. In M. Hersen & A. M. Gross (Eds.) *Handbook of clinical psychology: Vol I. Adults* (pp. 205–223). New York: Wiley.

**Websites**

StatSoft, Inc: http://www.statsoft.com

## Statistical Control

The control of nuisance variables via an experimental design or statistical technique is rooted in causal inference. To infer causality in a study, a researcher must be able to infer that the results are a result of the treatment and not unchecked nuisance variables. For example, in comparing different drugs to determine which one is most effective in reducing diastolic blood pressure, changes in diastolic blood pressure must be attributable to the drugs. Clearly, if one is not careful in the assignment of individuals to the different drug treatment conditions, then one could end up making the wrong causal inference because of group idiosyncrasies. To make a fair comparison of the drugs, one must have "an equal playing field." The researcher must control those variables that are known to affect diastolic blood pressure like age, health, physical activity, and diet. Because there is also a possibility of unknown variables affecting blood pressure, the researcher must be very careful. In most experimental situations, individuals are assigned at random to the treatment conditions to control for unknown as well as known nuisance variables that could upset the playing field. Ideally, participants in all the treatment conditions should be identical except for the drug being taken. Only under these conditions can one unmistakably infer causality. Using statistical procedures to adjust the result of a study for nuisance variable differences is the primary purpose of statistical control.

Using statistical procedures to adjust treatment comparisons for individual differences that could bias the comparison is especially important when randomization is not possible. Consider the following example, a researcher studying different methods of teaching math assigns a method to each classroom/teacher. In this situation, for many

reasons one cannot assign students to classrooms at random; that is, randomization is not possible. This situation is known as a field study. When randomization is not possible, one must seek other means of control. A way to try to assert control of nuisance variables is to use statistical techniques that equalize the playing field. Statistical techniques equalize the playing field by using the principle of conditionality. Consider age: If age plays a factor in blood pressure, it would be advantageous to use participants of the same age, say 40. Although possible, it is very unlikely to happen because it is very difficult to find only 40-year-old individuals to participate in the study. One can achieve a similar effect, however, with statistical control.

A plethora of statistical techniques can be used to attempt to achieve this "conditionalization" of nuisance variables. This entry discusses only the most fundamental ones: the analysis of variance with a blocked (subdivided) variable, the analysis of covariance, propensity scores, multiple regression, and the statistical control charts.

## Blocking

Sometimes, it is possible to equate nonequivalent groups by blocking a relevant variable. To be successful with this procedure, one must be able to identify and reliably measure the variables likely responsible for making the groups different and that are related to the experimental response *Y*. The variables must be measured before the treatment or they must not be reactive to the treatment. The identification of variables that are likely responsible for group differences is often not only a statistical question but also a theoretical one. Statistically, the primary criterion for selecting the blocking variable is a substantial correlation with the response variable. Go back to the blood pressure example; this time, assume the patients have been treated with drugs a, b, and c and that patients have not been assigned at random to the drugs. This comparison involves preexisting groups. If it is suspected by examining the groups and from a theoretical understanding of variables that can affect blood pressure that age is an important factor, then age should be incorporated into the analysis. One way of incorporating age into the analysis is by blocking age, that is, subdividing into different age categories. In this situation, the statistical procedure would be the analysis of variance (ANOVA) with a blocked factor to make the comparison. Table 1 illustrates the basic design.

**Table 1** Blocked Factorial Design

| *Age* | *Drug A* | *Drug B* | *Drug C* |
|---|---|---|---|
| 40–50 years | | | |
| 51–60 years | | | |
| 61–70 years | | | |
| Over 71 | | | |

By "equalizing" the rows in terms of age, or blocking, one can examine the relation between age and drug efficacy. The design controls identified individual difference by creating homogenous rows for the drug comparisons. It can be determined whether the drugs work differently across age groups; that is, whether an interaction exists between age and drug. A secondary effect of blocking is that reducing the variability in each cell increases the power of the statistical test. Clearly, a factorial design like this one could involve unequal *n*s and possibly an empty cell. Thus, one might or might not be able to estimate all of the cell means or test all three hypotheses: drugs, age, and age by drug. The factorial blocked design can be also used when randomization is possible.

## Analysis of Covariance

The analysis of covariance (ANCOVA) is similar to the block factorial ANOVA, but it does not group individuals into categories. The ANCOVA can be used with or without randomization. Instead of a blocking factor, ANCOVA relies on the statistical control of the covariate or concomitant variable. To control for age, as in the previous example, the ANCOVA model is given by

$$y_{i,j} = \mu + \alpha_j + \beta * (\text{age}) + e_{i,j}. \qquad (1)$$

The observation $y_{i,j}$ is a function of the drug effect $\alpha_j$ plus age and random component $e$. By including the quantitative covariate age in the model, the drug effect can be estimated without age. The drug effect is statistically adjusted for age differences among the groups. The coefficient $\beta$ represents the partial effect of age on response *Y*. Each mean for the *j* group is then given by

$$\mu_j = \mu + \alpha_j + \beta*(\overline{\text{age}}).$$

And the adjusted (for age) diastolic blood pressure mean is given by

$$\mu_j - \beta*(\overline{\text{age}}) = \mu + \alpha_j.$$

Next the adjusted means are compared to see the effect of the drugs on blood pressure. Note that the adjusted cell means are not affected by the average age in the group. It does not matter whether one group is older than the other, as long as there is no interaction between age and drug type. Furthermore, as in the blocked factorial design, the inclusion of the covariate in the model generally increases the power of the statistical test. If the relationship between age and $Y$ changes across groups, the model must be modified to include these differences,

$$y_{i,j} = \mu + \alpha_j + \beta_j*(\text{age}) + e_{i,j}. \qquad (2)$$

The model in Equation 2 indexes the beta to indicate that there is a beta per group. The model in Equation 2 can be compared against the simpler model in Equation 1 using an $F$ test to evaluate the advantage of adding the interaction. Additional covariates can also be added to the model if other variables are available and theoretically plausible.

As previously mentioned, an ANCOVA can be used in a randomized design or with intact groups. Under randomization, the primary use of ANCOVA is to reduce the noise or random error. With preexisting groups, however, group differences might be caused by real differences between the groups. The ANCOVA adjustment is a hypothetical one and reflects "how the data might have been, but not what they actually were." It behooves the researcher using ANCOVA with preexisting groups to be careful in the interpretation of the results. Also, one must weigh carefully the ANCOVA assumptions of linearity between $x$ and $y$, homogeneity of regression, and no measurement error on the covariate to make sure that they are met. Because preexisting (intact) groups have not benefited from randomization, unmeasured variables that are related to the response could pose a special threat to inference. Various researchers suggest that the best way to deal with unmeasured variables is to include numerous covariates in the model that are related to the outcome and that are not likely to be affected by the treatment. Multiple regression is a different version of the ANCOVA model that is used by researchers when the independent variables are all quantitative. Multiple regression will not be discussed here because of space constraints and because of its similarity to ANCOVA. Regardless of the technique used, theoretical justification should play a big role in the selection of the covariates to avoid artificial situations in statistical control. Albert Bandura, among others, has pointed out that in some cases, controlling a variable can lead to controlling a host of other variables that the researcher is not particularly interested in controlling, hence the importance of considering theory. For instance, controlling for grade point average (GPA) could also control for intelligence and self-esteem because of the correlation among the variables.

## Matching Subjects

To equalize two groups, subjects can be matched. In this technique, each subject in the control group is matched with a similar subject in the treatment group. Subjects without reasonable matches are left out of the comparison. The matching usually takes place along an important covariate, and it can take place before or after the experiment. One problem with matching, however, is that it is usually difficult to find good matches. If matching occurs after the experiment takes places, a covariate must be used that was not affected by the treatment. Investigators have cautioned about using matching across two different populations, because of the danger of ending up with matched samples that do not represent either population. For example, Figure 1 depicts two groups where the overlap between the two populations is marginal but the groups have been matched.

Matching these two groups to compare them would not yield a meaningful comparison, because the matched groups would not be representative of their populations.

## Propensity Scores

Statisticians have developed procedures for matching based on what has been called propensity scores. Propensity scores provide a way for adjusting for preexisting differences between two intact

groups, that is, groups that have not benefited from randomization. Propensity scores work by summarizing all the information about group differences from the covariates into a single variable, the propensity score. This is achieved by building *P*a statistical model to estimate the probability of group membership given the covariates, *P*(Group membership | covariates).

Subjects are then matched on their propensity scores across the groups; subjects with similar propensity scores are expected to have similar values on the background variables. Many techniques have been used to obtain propensity scores, including logistic regression and predictive discriminant analysis. As with all types of matching, it is important to ensure that a match is reasonably representative of both groups.

The importance of equalizing the groups in field studies can be illustrated with the following study of aspirin. Patricia Gum, Maran Thamilarasan, Junko Watanabe, Eugene Blackstone, and Michael Lauer conducted a study to evaluate the use of aspirin in coronary artery disease. In the study a group of subjects taking aspirin for approximately 3 years was compared with a group of subjects taking no aspirin. This was not a randomized clinical study, and researchers were comparing two intact groups. Of the 6,174 subjects in the study, 2,310 had been taking aspirin. The two groups were different on many covariates including gender, clinical histories, diabetes, and medication use. The simple univariate analysis comparing the two groups without any adjustment showed no association between aspirin use and cardiovascular mortality of 4.5% versus 4.5%. In fact, the two groups showed exactly the same mortality. However, once the groups were adjusted by matching subjects on their propensity scores they found group differences. The difference in mortality between the two matched groups was 4%. Aspirin users had a 4% lower risk of mortality than non-aspirin users (8%).

## Process Control Chart

Another important form of statistical control used by manufacturers, hospitals, and other organizations is the process control chart. Many versions of the control chart have been proposed. Many control charts have appeared in the literature including the Schewhart Chart, the Cumulative Sum Chart, and the Exponentially Weighted Moving Average Chart. All of these charts have one thing in common in that the *x*-axis is always time. In these charts, the researcher generally plots some sort of average against time. The charts also include upper and lower control limits. The range between the two control limits indicates the bound of "normal" variability. For example, a hospital can plot the mean monthly count of infections during the years to study infection rate across the years. These data are also used to compute the lower and upper control limits. Clearly, in a situation like this one, the upper control limit is more important. A dangerous or abnormal infection rate is one that goes above the upper control limit, signaling a cause for alarm. The basic premise of the control chart is to try to discern between random and not-random causes of variability. Going above the upper limit generally indicates to the staff that the increase in the infection rate is unlikely to be the result of random variation, a special cause. Of course, for control charts to be of value, they must be based on a substantial number of years. The control chart is also used in manufacturing to separate special from common (random) causes in variation. All processes have variability over time. The number of defects in a manufacturing process varies from one instance to the next. Manufacturers have learned that it is important to

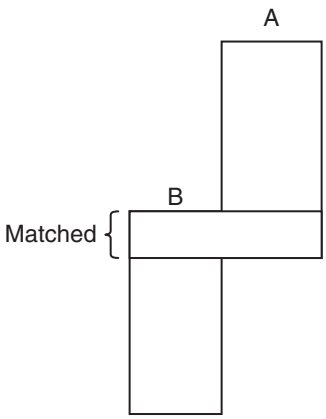



**Figure 1** Marginal Match

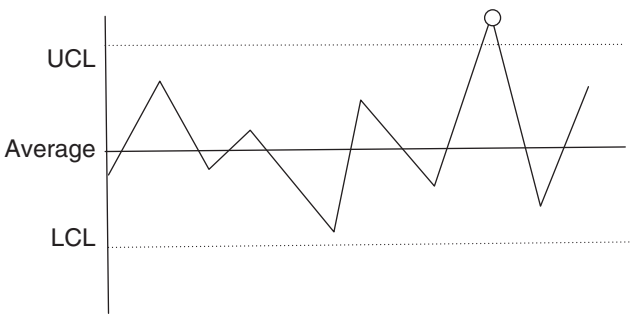



**Figure 2** Control Chart

understand what causes a large spike, a spike that goes outside of the limits, on the number of defects. Going outside the limits indicates that the process is outside of the expected variability and possibly in the range of a special cause. There is no space to discuss here how the upper and lower limits are obtained in a control chart, but it is tantamount of computing a confidence interval. When computing the limits and the process is not stable, it is important to try to determine and eliminate the causes of instability. If a special cause for the variability is found, the data are purged and the upper and lower limits recomputed. Once a stable process is identified, one that contains only random variability, the upper and lower limits are obtained. These limits in turn can be used in the chart to look for future abnormal situations. If a point is found outside the limits, then it is probable that this is an unusual situation that requires special consideration. Figure 2 depicts this situation.

### Additional Considerations

In many instances, a researcher can benefit from controlling unwanted sources of variability when making comparisons. Controlling for unwanted sources of variability should always go hand-in-hand with scientific theory. The failure to take into consideration theory when attempting to apply statistical control can lead the researcher to false conclusions. In contrast, especially with intact groups, failing to take preexisting group differences into consideration can also lead to the wrong inference.

*Jorge Mendoza and Susan Marcus-Mendoza*

***See also*** Analysis of Covariance (ANCOVA); Block Design; Matching; Multiple Regression; Propensity Score Analysis; Randomized Block Design

### Further Readings

Bandura, A. (1997). *Self-efficacy: The exercise of control.* New York: W. H. Freeman.

Breaugh, J. A. (2006). Rethinking the control nuisance variables in theory testing. *Journal of Business and Psychology, 20,* 429–443.

Cook, T. D. (1993). A quasi-sampling theory of the generalization of causal relationship. In L. B. Sechrest & A. G. Scott (Eds.), *New directions for program evaluation* (pp. 39–81). San Francisco: Jossey-Bass.

Gum, P. A., Thamilarasan, M., Watanabe, J. Blackstone, E. H., & Lauer, M. S. (2001). Aspirin use and all-cause mortality among patients being evaluated for known or suspected coronary artery disease: A propensity analysis. *Journal of the American Medical Association, 286,* 1187–1194.

Imbens, G. W., & Rubin, D. B. (1997). Bayesian inference for causal effects in randomized experiments with non-compliance. *Annals of Statistics, 25,* 305–327.

Rosenbaum, P. R. (1995). *Observational studies.* New York: Springer-Verlag.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70,* 41–55.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79,* 516–524.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66,* 688–701.

West, S. G., Biesanz, J. C., & Pitts, S. (2000). Causal inference and generalization in field settings: Experimental and quasi-experimental designs. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology.* Cambridge, UK: Cambridge University Press.

## *Statistical Power Analysis for the Behavioral Sciences*

Motivated by surveys indicating that data analyses published in top behavioral science and psychology journals between 1960 and 1982 were

seriously lacking in statistical power, Jacob Cohen wrote the book *Statistical Power Analysis for the Behavioral Sciences* to draw attention to the problem and to provide a handbook of statistical power analysis for behavioral scientists.

The book covers power analysis based on various statistical tests commonly used in behavioral science. These tests include the following: (a) the *t* test for means, (b) significance test of a product-moment correlation coefficient, (c) test of the differences between correlation coefficients, (d) test to determine whether a proportion is 0.5 and the sign test, (e) test of the differences between proportions, (f) chi-square tests for goodness of fit and contingency tables, (g) analysis of variance (ANOVA) and covariance, (h) multiple regression and correlation analysis, and (i) a multivariate version of the multiple regression and correlation analysis.

Each test is discussed in a separate chapter, which starts with a description of the test, continues with a definition of the effect size index, and ends with a power table and a sample size table for the test, along with (for most tests) an explanation of how to use the power table for significance testing. Using the tables provided in the book, one can easily find the power and necessary sample size by locating the specified parameters. For example, given an effect size and a sample size for a *t* test for between-group means, one can find the power by locating the specified effect and sample sizes. Similarly, one can find the required sample size when the effect size and power are specified. In most power tables, a significance criterion in the form of the sample effect size is given to facilitate significance testing if data are available.

The book also discusses issues related to (a) understanding the magnitude of effect sizes, (b) the role of psychometric reliability, and (c) the efficacy of "qualifying" (differencing and partialling) dependent variables in the context of power analysis. A description of the computational procedures associated with each test is provided at the end of the book.

This book is useful in that it not only provides results for power and sample size calculations based on different tests but also describes the tests and the corresponding computational procedures so that interested readers can reproduce the results, presumably with additional computing aid. This book does have limitations. The first edition of the book was published in 1969. Despite the addition of new material in the revised and second editions (published in 1977 and 1988, respectively), descriptions of power analysis for some important statistical methods currently in common use in behavioral research are missing. These methods include logistic regression analysis, repeated measures ANOVA, and McNemar's test for the difference between correlated proportions, among others. Thus, researchers might prefer to use one of the various software packages currently available that will conduct power and sample size calculations. Such software is typically user friendly and quickly generates results with complementary tables and/or graphics (e.g., provides power statements with visual demonstrations such as plots of power curves). In addition, current software packages incorporate power analysis for most of the recently developed statistical procedures that are not covered in the book. Individuals interested in understanding power and sample size calculations based on standard and conventional statistical tests might find this book useful. If instead the main goal is to calculate power and/or sample size, especially using modern advanced statistical techniques, then software might prove to be more convenient. Popular software of this sort includes nQuery, Power and Sample Size (PASS), and others.

*Yisheng Li*

***See also*** Clinical Significance; Nonsignificance; Power; Significance, Statistical; Significance Level, Concept of; Significance Level, Interpretation and Construction; Type II Error

**Further Readings**

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences* (1st ed.). Hillsdale, NJ: Lawrence Erlbaum.

## Stepwise Regression

Stepwise, also called stagewise, methods in fitting regression models have been extensively studied

and applied in the past 50 years, and they still remain an active area of research. In many study designs, one has a large number $K$ of input variables and the number $n$ of input–output observations $(x_{i1}, \ldots, x_{iK}, y_i)$, $1 \leq i \leq n$, is often of the same or smaller order of magnitude than $K$. Examples include gene expression studies, where the number $K$ of genomic locations is typically larger than the number $n$ of subjects, and signal or image reconstruction, where the number $K$ of basis functions to be considered exceeds the number $n$ of measurements or pixels. Stepwise methods are perhaps the only computationally feasible ways to tackle these problems, and certain versions of these methods have recently been shown to have many desirable statistical properties as well.

Stepwise regression basically carries out two tasks sequentially to fit a regression model

$$y_i = \beta^T x_i + \varepsilon_i, 1 \leq i \leq n, \qquad (1)$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)^T$ is a vector of regression parameters, $\boldsymbol{x_i} = (x_{i1}, \ldots, x_{iK})^T$ is a vector of regressors (input variables), $\varepsilon_i$ represents unobservable noise, and $y_i$ is the observed output. The first task is to choose regressors sequentially and the second task is to refit the regression model by least squares after a regressor has been added to the model. For notational simplicity, assume that the $y_i$ and $x_{ij}$ in Equation 1 have been centered at their sample means so that Equation 1 does not have an intercept term.

To begin, stepwise regression chooses the regressor that is most correlated to the output variable (i.e., such that the [sample] correlation coefficient between $y_i$ and $x_{ij}$ is the largest among the $K$ regressors). One then performs least squares regression of $y_i$ on the selected regressor $x_{ij}$, yielding the least squares fit $\widehat{y}_i = \widehat{b}_j x_{ij}$ and the residuals $y_i - \widehat{y}_i$. A variable selection criterion is then applied to determine whether the chosen regressor should indeed be included. If the criterion accepts the chosen regressor, then the researcher repeats the stepwise procedure to the remaining regressors but with $y_i - \widehat{y}_i$ in place of $y_i$. More generally, after the regressors labeled $j_1, \ldots, j_k$ have been included in the model and the residuals $e_i = y_i - (\widehat{b}_{j1} x_{i,j1} + \cdots + \widehat{b}_{jk} x_{i,jk})$ have been computed, the researcher chooses $\tilde{j} \notin \{j_1, \ldots, j_k\}$ such that the correlation coefficient between $e_i$ and $x_{i,\tilde{j}}$ is the largest among the remaining $K - k$ input variables, and it performs least squares regression of $e_i$ on $x_{i,\tilde{j}}$, yielding a new set of residuals $\widetilde{e}_i$, which are used in the criterion to determine whether the regressor labeled $\tilde{j}$ should be included. If the criterion rejects the regressor, then it is not included in the model, and the stepwise regression procedure terminates with the set of input variables $x_{i,j1}, \ldots, x_{i,jk}$.

A traditional variable selection criterion is based on the $F$ test of $H_0 : \beta_j = 0$ in the regression model $y_i = \beta_{j1} x_{i,j1} + \cdots + \beta_{jk} x_{i,jk} + \beta_{\tilde{j}} x_{i,\tilde{j}} + \varepsilon_i$ to determine whether the regressor labeled $\tilde{j}$ should be added to the model that already contains the regressors labeled $j_1, \ldots, j_k$. If the $F$ test rejects $H_0$ at significance level $\alpha$, which is often chosen to be 5%, then the regressor labeled $j$ is included in the model. Otherwise, $\beta_j$ is deemed to be not significantly different from 0, and therefore the corresponding regressor is excluded from the model. Note that because this test-based procedure carries out a sequence of $F$ tests, the overall significance level can differ substantially from $\alpha$. Such an $F$ test of whether a particular regressor in a larger set of input variables has regression coefficient 0 is called a *partial* F *test,* and the corresponding test statistic is called a *partial* F *statistic.*

The forward selection of variables presented previously, until the latest entered variable is not significantly different from 0, is often combined with a backward elimination procedure that begins by computing the partial $F$ statistic for each included regressor. The smallest partial $F$ statistic $F_L$ is compared with a prespecified cutoff value $F^*$ associated with the partial $F$ test of $H_0 : \beta_L = 0$. If $F_L > F^*$, the researcher terminates the elimination procedure and includes all the entered variables. If $F_L < F^*$, then the researcher concludes that $\beta_L$ is not significantly different from 0 and removes its associated input variable from the regressor set. This backward elimination procedure is then applied inductively to the set of remaining input variables at every stage.

Although this test-based method using partial $F$ tests was the earliest variable selection method and is still widely used in software packages, a better approach is to choose models based on their

performance assessment, which relates to how well a model can predict the outputs of future test data generated from the actual regression model. Without actually generating future test data, in-sample estimates of out-of-sample performance have been developed for the linear regression model in Equation 1. These include Mallows' $C_p$ statistic, Akaike's information criterion AIC($p$), and Schwarz's Bayesian information criterion BIC($p$), for a set of $p$ input variables included in the regression model. Let $\mathrm{RSS}_p = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ denote the residual sum of squares after performing least squares regression of $y_i$ on these $p$ regressors, and let $\sigma_p^2 = \mathrm{RSS}_p/n$, $s_K^2 = \mathrm{RSS}_K/n$, recalling that $K(\geq p)$ is the number of available regressors. Then $C_p$, AIC($p$), and BIC($p$) are defined by

$$C_p = (\mathrm{RSS}_p/s_K^2) + 2p - n, \tag{2}$$

$$\mathrm{AIC}(p) = \log(\hat{\sigma}_p^2) + 2p/n, \tag{3}$$

$$\mathrm{BIC}(p) = \log(\hat{\sigma}_p^2) + (p \log n)/n. \tag{4}$$

Because each quantity in the previous example is an in-sample estimate of some expected prediction loss, the criterion-based approach is to select the model with the smallest value of Equations 2, 3, or 4.

The direct application of a variable selection criterion to the set of all possible subsets of $K$ variables leads to a large combinational optimization problem that is computationally expensive for large $K$. To circumvent these difficulties, one can use stepwise methods consisting of forward addition of variables followed by backward elimination of the type described previously but with the partial $F$ statistics replaced by changes in the criterion between the models with $p$ and with $p+1$ regressors. Although the theory underlying the selection criteria in Equation 2 through Equation 4 assumes $K$ to be substantially smaller than $n$, recent advances in the literature have shown how Equations 3 and 4 can be modified for the case where $K$ is larger than $n$, and have also developed efficient computational methods to perform stagewise regression in this case.

*Tze Leung Lai and Ching-Kang Ing*

***See also*** *F* Test; Least Squares, Methods of; Multiple Regression; Partial Correlation; Semipartial Correlation Coefficient

**Further Readings**

Akaike, H. (1973). *Information theory and an extension of the maximum likelihood principle*. Presented at the second international symposium on Information Theory. Budapest, Hungary: Akademiai Kiado.

Bühlmann, P. (2006). Boosting for high-dimensional linear models. *Annals of Statistics, 34,* 559–583.

Hastie, T., Tibshriani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer-Verlag.

Ing, C. K., Lai, T. L., & Chen, Z. (2009). *$L_2$-boosting and consistent model selection for high-dimensional sparse linear models* (Technical Report). Palo Alto, CA: Stanford University.

Mallows, C. L. (1973). Some comments on $C_p$. *Technometrics, 15,* 661–675.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461–464.

# Stratified Sampling

Stratified random sampling (usually referred to simply as *stratified sampling*) is a type of probability sampling that allows researchers to improve precision (reduce error) relative to simple random sampling (SRS). The population is divided into non-overlapping groups, or *strata,* along a relevant dimension such as gender, ethnicity, political affiliation, and so on. The researcher then collects a random sample of population members from within each stratum. This technique ensures that observations from all relevant strata are included in the sample.

## Types of Stratified Sampling

### *Proportionate Stratification*

A sample with proportionate stratification is chosen such that the distribution of observations in each stratum of the sample is the same as the distribution of observations in each stratum within the population. The *sampling fraction,* which refers to the size of the sample stratum divided by the size of the population stratum ($n_h/N_h$), is

equivalent for all strata. Proportionate stratification is primarily useful when a researcher wishes to estimate values in a population and believes that different groups within the population differ on the variable of interest. For example, suppose a university administrator wants to measure attitudes about imposing a new student fee among students at a university with 10,000 students, including 4,500 in the humanities, 3,000 in the sciences, and 2,500 in business. If the administrator wanted a sample of 200 (2% of the students), she would randomly sample 90 humanities majors (2% of 4,500), 60 science majors (2% of 3,000), and 50 business majors (2% of 2,500). It is essential that the stratification factor be related to the variable of interest. Stratification by major will improve the precision of the estimate of attitude in the population to the extent that students in different majors have different attitudes about the fee increase and relatively similar attitudes to other students within their major.

### *Disproportionate Stratification*

In disproportionate stratification, the sampling fraction is not the same across all strata, and some strata will be oversampled relative to others. Disproportionate stratification is primarily useful when a researcher wants to make comparisons among different strata that are not equally represented in the population. For example, a researcher in the United States who is interested in comparing learning outcomes among children of different ethnic backgrounds might use disproportionate stratification to ensure that the sample includes an equal (or at least sufficient) number of children from minority groups. Suppose that a school district of 10,000 children is 80% Caucasian (8,000 children), 15% African American (1,500 children), and 5% Asian American (500 children). If a researcher draws a simple random sample of 100 children (1% of the population), the sample is likely to contain about 15 African American and 5 Asian American children, but by random chance those groups could be considerably smaller. To ensure adequate samples of African American and Asian American children, the researcher can increase the sampling fraction of these subgroups. Say the researcher decides to include 40 students from each group in the sample. In this case, the sampling fraction for Caucasians is 40/8,000 = .005, for African Americans, it is 40/1,500 = .03, and for Asian Americans it is 40/500 = .08.

Although a disproportionate stratified sample is optimal for making comparisons among strata, it is not ideal for making estimates regarding the total population, because the sample does not evenly represent the population. If the researcher wishes to draw conclusions about the population as a whole based on a disproportionate stratified sample, then results must be weighted to reflect the different sampling fraction of each stratum.

## Benefits of Stratified Sampling

### *Increasing Representativeness of the Sample*

In a simple random sample, every member of the population has an equal chance to be selected as a member of the sample. In theory, this should produce samples that are perfectly representative of the population. In practice, this is often not the case, particularly when the sample size is small or when the population contains small subgroups that might be underrepresented or missed entirely in a randomly drawn sample. Stratified sampling helps ensure that all relevant groups are represented in the final sample.

### *Improving Precision of Estimation*

Proportionate stratified sampling can provide an increase in precision (meaning a reduction in error) relative to simple random sampling, if variability within each stratum is smaller than variability across the entire population. Variance between strata is removed from the total variance, thus reducing the standard error and improving precision of the estimate. Error variance will be reduced to the extent that groups are homogeneous within strata and heterogeneous between strata. Thus, it is important to choose stratification variables that divide the population into groups that are distinct from each other and consistent within each group.

If the strata do not differ (the between-strata variance is zero), then the total variance will not be reduced; however, the total variance will never be increased by using proportionate stratification.

### *Economy*

When used properly, stratified sampling can substantially reduce the number of data points necessary to achieve precise population estimates and statistically significant comparisons. A researcher can achieve the same degree of power using a smaller sample if the within-strata variability is smaller than the total population variability. In addition, stratified sampling can be used to reduce overall costs of data collection through effective allocation of resources. For example, when predicting the outcome of a presidential election, it is necessary to stratify by state because the general election is decided by the popular vote within each state rather than by the overall popular vote in the nation. In states where the voting trends are very strong, such as Oklahoma, which voted 66% for McCain versus 34% for Obama in the 2008 Presidential election, accuracy in predicting the election outcome can be achieved with a relatively smaller sample compared with states where the election is very close. Small fluctuations in sampling are unlikely to lead to an incorrect prediction in a population with such high internal consistency. In contrast, accurate prediction requires a larger sample in states with close contests, such as North Carolina, which elected Obama with a final advantage of fewer than 14,000 votes out of more than 4 million votes cast. Knowing which states can be predicted with smaller samples can allow pollsters to distribute their resources more efficiently.

### *Reducing Bias Caused by Nonresponse*

Biases caused by nonresponse might be reduced, although not eliminated, using stratification. When individuals who are initially selected for a study choose not to participate, this leads to a biased sample, particularly if certain groups of people are more likely than others to eschew participation. Using stratified sampling during the data-collection phase, nonresponders can be replaced by participants from the same stratum, thus reducing the bias. After data collection, it is possible to use stratification weights to correct for minor sample biases resulting from nonresponding; this procedure is known as *poststratification*.

## Concerns About Stratified Sampling

### *Not Knowing Distribution of Strata in the Population Leads to Inaccurate Estimates*

A properly stratified sample can be difficult to achieve. It is essential to have up-to-date and accurate information about the strata in the population. If the distribution of strata in the population is measured or estimated inaccurately, then the stratified sample will mimic this inaccuracy and will not be representative of the population. For example, using census data to determine strata might lead to inaccurate stratification if the distribution of population characteristics has changed during the time since the census was conducted. When disproportionate stratification is used, inaccurate knowledge about population strata will lead to incorrect weighting, which leads to inaccurate population estimates based on the sample.

In some cases, it is not possible to have complete knowledge about population strata prior to data collection. For example, election pollsters cannot know in advance which people will actually vote in the election and must do their best to estimate the voter turnout among different groups. Stratifying by gender and ethnicity according to their overall population distributions will lead to inaccurate election predictions if voter turnout is different among men and women from different ethnic groups.

### *Stratifying on an Irrelevant Factor Does Not Improve Precision*

Stratified sampling will confer no benefit if the characteristics used to divide strata are not relevant to the research question(s). For example, if student attitudes about a university fee increase do not vary across different majors, then stratifying the sample by major will be a wasted effort. In studies that measure multiple variables, relevant stratification factors might not be the same for all variables. When using a complex survey, it might not be practical or even possible to stratify along all dimensions that are relevant to the variables. Researchers must balance the cost of creating a stratified sample against the measurement precision gained by stratification.

### *Disproportionate Stratification Can Lead to Reduced Precision*

Population estimates made using a proportionate stratified sample will never be less precise compared with estimates made using an SRS of the same size. However, a disproportionate stratified sample might be less precise compared with SRS. This problem occurs when strata that are oversampled relative to their population distribution have higher variance compared with strata that are undersampled. Disproportionate stratified sampling can lead to reduced overall precision, if the variance in oversampled strata is greater than the variance in the overall population.

### *Stratification Does Not Eliminate the Need for Probability Sampling Techniques*

Stratified sampling requires that samples be drawn randomly within each stratum. Using nonprobability techniques (e.g., volunteers) to sample within strata is called *quota sampling*. Quota sampling helps ensure that members of relevant groups are included in the sample, but it does not ensure that the sample is representative. A convenience sample is still a convenience sample, even if the researcher has taken care to recruit volunteers from different genders, education levels, and so on.

## Analytic Strategies With Stratified Samples

Estimating population values using a stratified sample is more complicated compared with using SRS. The analysis effectively involves two steps. First, sample statistics are calculated for each stratum, and then these estimates are combined using weights that reflect the relative distributions of the strata in the population. As shown in the following formula (from Richard Scheaffer, William Mendenhall, and Lyman Ott) for sample mean based on a stratified sample ($\bar{y}_{st}$) with $L$ strata, the sample mean is computed for each stratum ($\bar{y}_i$), multiplied by the population size of that stratum ($N_i$), and summed across all $L$ strata before dividing by the total population size ($N$).

$$\bar{y}_{st} = \frac{1}{N}\sum_{i=1}^{L} N_i \bar{y}_i.$$

The error variance ($s^2_{\bar{y}_{st}}$) is computed as follows, where $s_i^2$ is the variance of a given stratum and $n_i$ is the sample size for a given stratum:

$$s^2_{\bar{y}_{st}} = \frac{1}{N^2}\sum_{i=1}^{L} N_i^2 \left(\frac{N_i - n_i}{N_i}\right)\left(\frac{s_i^2}{n_i}\right).$$

Statistical analysis programs such as SPSS , an IBM company, formerly called PASW® Statistics, provide commands to specify variables used for stratification and compute parameter estimates based on the stratified sample.

*Kristi M. Lemm*

***See also*** Probability Sampling; Proportional Sampling; Quota Sampling; Random Sampling; Sampling

**Further Readings**

Kalton, G. (1983). *Introduction to survey sampling*. Beverly Hills, CA: Sage.

Lee, E. S., & Forthofer, R. N. (2006). *Analyzing complex survey data* (2nd ed.). Thousand Oaks, CA: Sage.

Parten, M. (1966). *Surveys, polls, and samples: Practical procedures*. New York: Cooper Square Publishers.

Scheaffer. R. L., Mendenhall, W., III, & Ott, R. L. (2006). *Elementary survey sampling* (6th ed.). Belmont, CA: Thomson Brooks/Cole.

Sudman, S. (1983). Applied sampling. In P. H. Rossi, J. D. Wright, & A. B. Anderson (Eds.), *Handbook of survey research* (pp. 145–194). New York: Academic Press.

Thompson, S. K. (2002). *Sampling* (2nd ed.). New York: Wiley.

# Structural Equation Modeling

Structural equation modeling (SEM) belongs to the class of statistical analyses that examines the relations among multiple variables (both exogenous and endogenous). The methodology can be viewed as a combination of three statistical techniques: multiple regression, path analysis, and factor analysis. It has the purpose of determining the extent to which a proposed theoretical model, which is often expressed by a set of relations among different constructs, is supported by the collected data. SEM is, therefore, a technique for

confirmatory instead of exploratory analysis. This entry represents a nonmathematical introduction to SEM, with an emphasis on its benefits, usage, and basic underlying assumptions. Principal concepts related to the methodology and summarized comparisons with other multivariate analyses are also presented. Then, the entry outlines and discusses a pragmatic six-step framework to carry out the SEM statistical analysis. Advances in SEM with respect to the relaxation of its fundamental assumptions conclude the entry.

## Principal Concepts

Essentially, SEM statistically tests hypotheses of multiple relations among measured variables. As depicted in Figure 1, the conceptual variables or constructs are denominated latent variables. Hence, SEM models these qualitative relations among the latent variables and then tests the extent to which these relations explain the data according to a fit criterion. In many cases, the latent variables cannot be directly observed and are further specified and estimated through the measurements of underlying variables (the *X*s and *Y*s in Figure 1). Besides, through the model's latent variables and their underlying measured variables, SEM can explicitly account for measurement errors.

In general, two distinct but interconnected models lie within a single SEM analysis: a measurement model, aimed to evaluate the extent to which the measured (underlying) variables effectively capture the concepts (or constructs) in the model; and a structural model, aimed to simultaneously test all the causal relations (i.e., the links among the constructs) found in the overall SEM model. It should be noted that the consistency of the causal relations among the latent variables does not rely on the confirmatory testing but on the theoretical support of the model proposed.

SEM has its origins from several different multivariate methods. On one side, the measurement model originates from a much broader category of factor analysis. This statistical technique is often used to study patterns within a set of variables, with the objective to identify their underlying structure. These groupings of variables are known as factors. Initial development of factor analysis came from Charles E. Spearman in 1904 by his work on the identification of underlying structure of mental abilities.

Factor analysis is generally considered an exploratory technique. On the contrary, confirmatory factor analysis (CFA) requires a priori assumptions about the structure of the variables and serves the a purpose similar to that of the measurement model in SEM framework.

The second constituent, path analysis, corresponds to the structural model within SEM analysis. It allows a researcher to model complex relationships among various variables and visually represents them in a path diagram. Path analysis compares the regression weights obtained from a proposed causal model to the correlations obtained from the data, and estimates the fit of the data to the proposed model.

In contrast to SEM, path analysis does not perform the simultaneous estimation of all causal relations specified in the model. Path analysis was first proposed by Sewall G. Wright in 1918 on the confirmation of the effect of factors (the causes) leading to animal bones' dimensions (the effects).

Third, SEM relates to multiple regression. This technique is used to learn about the relation between several independent variables and a dependent variable. Instead of dependent and independent variables, SEM identifies and uses exogenous and endogenous latent variables. Exogenous latent variables, on the one hand, are mostly equivalent to the independent variables in multiple regression because they are not determined by other latent variables in the model. Endogenous latent variables, on the other hand, are dependent on other latent variables within the SEM structural model. In contrast to multivariate regression analysis, SEM allows an endogenous variable to serve as both dependent and independent constructs at the same time along with simultaneous estimation of all model equations (i.e., relations in the model). In Figure 1, the latent variables one, two, and four are exogenous, and the latent variables three and five are endogenous.

Multiple regression establishes correlational links between independent variables and a dependent variable. SEM attends to three different types of links between the variables: (1) the relation between the latent variable and the measured variable(s), (2) dependent relations between latent variables, one exogenous or endogenous latent variable—the cause—and one endogenous

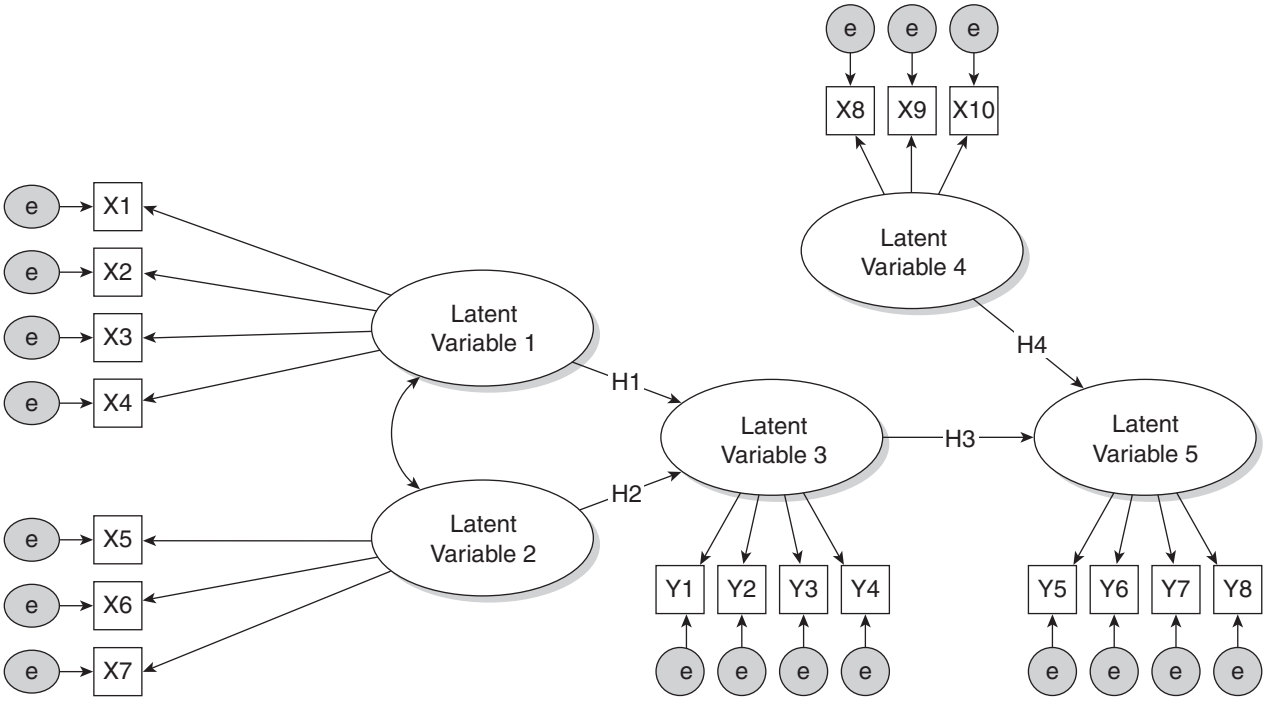



**Figure 1** An Example of SEM Involving Measurement and Structural Models

latent variable—the effect, and (3) correlational relations among exogenous latent variables, in which no causality is presumed. The three types of links are represented in Figure 1.

The modern formulation of SEM was developed by numerous statisticians and methodologists, among them include Michael W. Browne, Karl G. Jöreskog, Dag Sorbom, and Bengt O. Muthén.

SEM is used widely in scientific research because it offers clear advantages over other multivariate techniques. SEM is also the analytical technique of preference in nonexperimental research designs (where randomization and manipulation are not possible). Many researchers realize its great advantage of being able to test multiple relations and to account for multiple variables without compromising statistical power. At the same time, SEM integrates the test of the proposed theoretical model and the study of the errors of measurement in a comprehensive approach.

Current commercial packages for conducting SEM have simple interfaces and are easy to use. Among them, LISREL, AMOS (SPSS interface; SPSS is an IBM company formerly called PASW® Statistics), Mplus, and EQS are the packages of preference.

## Six-Step Framework

SEM practice requires attention to both the theoretical support of the model and the experimental design. There are many basic steps, as summarized in Figure 2, to perform a typical SEM analysis. Initially, the constructs are defined and the models involving the underlying variables are specified, paying special attention to the measurement model and the inherent experimental design. After the estimation of the path coefficients in the model, the collected data are used to determine the fit of the model. Once validity of the latent variables has been confirmed, the researcher should focus on the structural model. In an iterative process, the researcher can test structural variations of the model to assess their fit to the data. A final interpretation of the results should be based on the fit of model, the design of study, and the extent of how closely the gathered data conforms to the basic SEM assumptions.

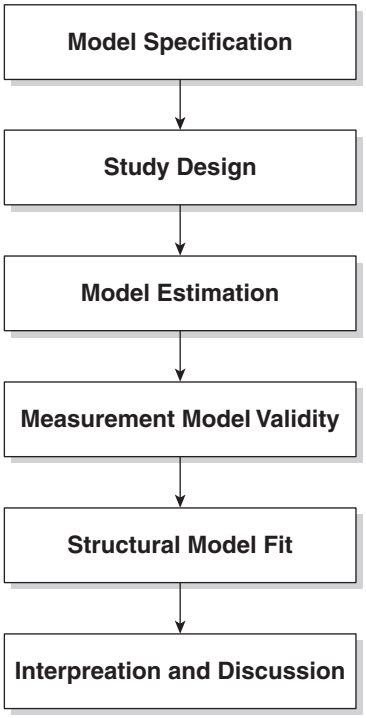



**Figure 2** Six Steps in a Typical SEM Framework

### *Step One: Model Specification*

The initial step in the development and testing of a SEM model focuses on identifying the definitions and operationalization of latent variables and forming the basis for the measurement model. The appropriate measurement of each latent variable through the use of indicators (i.e., underlying variables Xs and Ys in Figure 1) that properly represent the true subdimensions of the construct is essential not only to draw valid and meaningful conclusions but also to maintain the integrity of the overall SEM analysis.

Likewise, a solid theoretical support is also required to generate the initial structural model, which connects latent variables (constructs) through both causal and correlational links. Causality implies a directional relation between a construct identified as the cause and another construct identified as the effect. John Stuart Mill has proposed the following three requirements to demonstrate the existence of a causal relation: (1) the cause must precede the effect, (2) the cause is nonspuriously related to the effect, and (3) there is no plausible explanation for the effect other than the cause. To justify the validity of the conclusions drawn from the SEM analysis, a researcher has to attend to previous findings to the greatest extent possible. The formal process of specifying the structural model consists of deciding which path parameters are fixed to zero (for those constructs without any assumed causality) and which parameters are free and can thus be estimated and tested by the data. The number of free parameters used in the structural model also has implications for the design of the study. More details are discussed in the next section.

Different approaches are available for modeling development and testing, depending on the stage in the scientific discovery process the researcher is at and the previous/existing knowledge on the links found in the model. With the primary purpose of conducting confirmatory analysis, a single model can be proposed and tested against the data. Otherwise, alternative models can be generated by the existing theories and compared using their goodness of fit to the data. To refine a model, one can adopt a third strategy that progressively modifies the experimental model by accounting for the improvement in its ability to fit the data.

Despite the strategy followed and the final level of fit achieved, a researcher must keep in mind that the tested model holds only for the sample included in the study. Any other causal generalizations with respect to persons, settings, treatments, and outcomes must be carefully considered and evaluated based on existing conceptual frameworks and theories.

SEM allows a researcher to test theoretical propositions using nonexperimental data, resulting in valid conclusions only when theoretical rationale is consistent. Furthermore, it is possible that a large number of latent variables can lead to multiple alternative models with similar levels of fit to the empirical data. A researcher must use his or her own knowledge and judgment to develop and obtain a valid and meaningful model. In summary, a priori specifications are essential to SEM analysis. This precursory requirement characterizes SEM as a confirmatory (rather than an exploratory) statistical technique.

### *Step Two: Study Design*

After the theoretically constructed model has been specified, several considerations affecting the

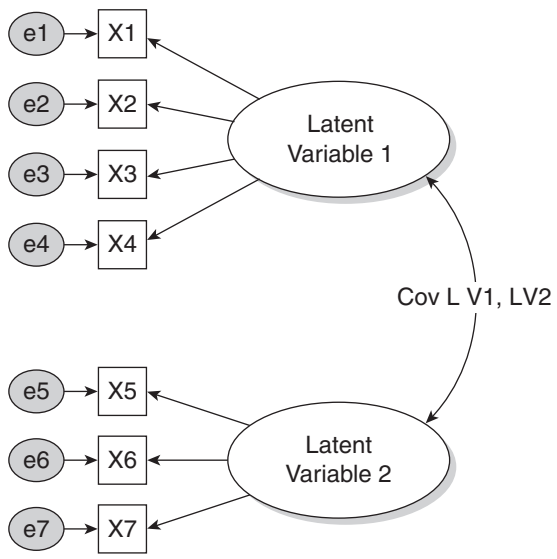



**Figure 3** Measurement Model of SEM

design of the study must be made. Each latent variable should be specified through the measurement of multiple indicator variables (denoted by X1 to X4 in Figure 3). These indicator variables would be operationalized as scale items in modeling. Although it is recommended to use multiple indicator variables per latent construct and to adopt previously tested items for better theoretical support, it is up to the researcher to decide how many indicators are to be included in a given latent construct. At the same time, the researcher should also be aware that the type of scale selected for each item (nominal, ordinal, interval, or ratio) and the range allowed for the measured values (e.g., three-point Likert scale versus seven-point Likert scale) do exert an influence on the covariances among the latent variables, which in turn affect the final outcomes and the goodness of model fit.

In addition, there are two other important factors for design of study—the sampling method and the sample size. The sample used in SEM should be representative of the population to derive useful causal generalizations. The sampling procedure should also allow room or flexibility regarding missing data points. Generally speaking, the minimum sample size required is 200. Very large samples, especially the ones with more than 400 records, might result in data overfitting and poor performance and generalizability of the SEM model. The number of free parameters in the model, which increases with the number of indicator variables and the number of latent variables, should be considered to determine the sample size. Other factors affecting the required sample size are associated with the statistical assumptions of SEM, namely, deviations from multivariate normality (critical when maximum-likelihood estimator [MLE] is used to estimate the path coefficients), missing data, and specification errors (omitted variables), which might distort the correlations between the error terms and the exogenous variables.

Finally, although it is not desirable to have missing data points, the situation might inevitably occur as a result of uncontrollable factors. In such case, a researcher should examine the missing samples in the data set as a precautionary step. The general rule of thumb is that the samples should be missing completely at random (MCAR), that the missing data do not depend on another variables, or that samples are missing at random (MAR) when the missing data for a variable are related to another variable(s). Missing data should never represent more than 10% of the total data and constitute any explicit patterns.

### *Step Three: Model Estimation*

The model estimation step follows the assessment of the free parameters identified in step one. Unless they are fixed, links in the SEM model, including the ones between the error terms and the indicator variables and the causalities in structural model (see Figure 1), need to be estimated. Simply put, after compiling the data with respect to the model specification(s) determined in step two, sample covariance (or correlation) matrix for various entities can be calculated. Then, in the model estimation process, each parameter is estimated in a way that it reassembles the sample covariance (or correlation) matrix to the maximum extent. Mathematically, a fitting function is used to minimize the difference between the observed and the estimated matrices. The most common fitting function is based on MLE, although other methods such as generalized least squares (GLS), weighted least squares (WLS), and asymptotically distribution free (ADF) estimation are frequently available in the software packages. In the situation where MLE shows great sensitivity to non-normality or the multivariate normality assumption is not met,

**Table 1** Goodness-of-Fit Measures

| *Goodness-of-Fit Measures* | *Statistical Test* | *Advantages* |
|---|---|---|
| Chi-Square ($\chi^2$) | Yes | Sensitive to sample size and the number of parameters |
| Adjusted Goodness-of-Fit lndex (AGFI) | No | Adjusted for the complexity of the model and has lower sensitivity to sample size than $\chi^2$ |
| Root Mean Square Error (RMSE) | No (confidence interval can be reported) | Corrects for sample size and model complexity |
| Comparative Fit lndex (CFI) | No | Partly insensitive to sample size. Allows incremental verification of fit |
| Wald Test | Yes | Can evaluate difference in $\chi^2$ when a parameter is added or dropped |

the other fitting functions, especially ADF estimation, are preferred.

It is worthwhile to note that, although it is recommended to compare covariance matrices in SEM analysis, the fit might also be assessed using correlation matrices. A correlation matrix is standardized, dimensionless, and easier to understand; however, it offers less information than the covariance matrix. A researcher should use his or her best judgment to determine which option better accomplishes the goals of the study. Some researchers have suggested using multistep approaches in which the estimation of the parameters of the measurement model and the estimation of the parameters of the structural model are performed in different steps.

### *Step Four: Measurement Model Validation*

Validity refers to the extent in which an inference is truthful and consistent with previous empirical findings and existing knowledge. The measurement model validation step explores and verifies the validity of the tested model through an examination of the validity of the measurement model's latent variables, that is, an assessment of the goodness of fit of the specified model.

The validity of the latent variables (or constructs) should be examined from both statistical and conceptual perspectives. Their assessment is a process covering the context of appropriateness of construct definitions, correct identification of internal structure (content validity, face validity), adequate operationalization (or implementation) through multiple items—indicator variables (predictive, concurrent, convergent, and discriminant validities should be evaluated), and estimation of reliability of the measures.

The estimation of a goodness of fit is based on a statistical evaluation of the similarity between the actual covariance matrix and the estimated covariance matrix. A larger degree of similarity between the two matrices usually implies a better fit of the model to the data. Nevertheless, researchers should be cautious of the problem of overfitting in SEM analysis.

To assess the fit, several measures of fit can be considered. The most common measure of fit, chi-square ($\chi^2$), statistically assesses the significance of the difference between the observed sample covariance matrix and the estimated covariance matrix. Alternative measures of fit, along with some brief comments, are compared with $\chi^2$ in Table 1. The number of parameters and parameters to be estimated (calculated and referred as degrees of freedom) is frequently used to account and compensate for model complexity. The final adoption of a given model should be considered on a case-by-case basis, and special attention should be paid to the characteristics of the sample data, the proposed model, and the overall design of the study. It is recommended to use several fit measures when assessing the goodness of fit of the measurement model.

It should be pointed out that validity issues in the measurement model should be corrected before advancing to the estimation of the structural model.

### *Step Five: Structural Model Estimation*

The estimation of the structural model relies on the previous specification of the relations in the

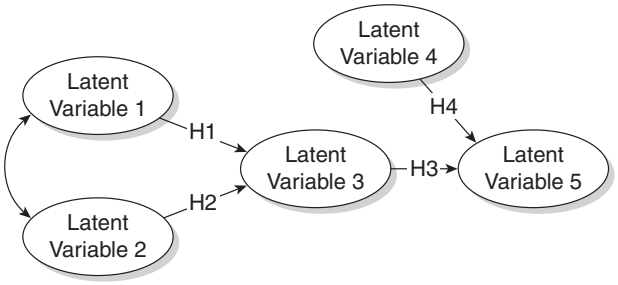



**Figure 4** An Example of a Structural Model in SEM

model (step one). Figure 4 presents an example of a structural model with four hypotheses.

To assess the validity of the structural model, a structural goodness of fit is calculated. Similar to step four, the estimated covariance (or correlation) matrix is compared with the covariance (or correlation) matrix calculated from the observed data. The $\chi^2$ statistic is a general guideline to establish the fit of the model. A specific hypothesis relating two of the latent variables in the model can be evaluated by looking at the significance and value (to assess the directionality) of the path parameters. Corresponding to the model selection and refinement strategies described in step one, competing models can be compared using AGFI or CFI (described in Table 1) or their $\chi^2$ values tested for significance using the chi-square difference test statistic ($\Delta\chi^2$).

The estimation of the structural model is generally conducted in an iterative manner, in which the researcher modifies the structure of the model and reestimates until a satisfactory goodness of fit is achieved.

### *Step Six: Interpretation and Discussion*

Virtually essential in any scientific endeavor, the final step in the SEM analytical framework is to communicate and interpret the results of the study efficiently. At this point, the researcher should recapitulate the characteristics and consistency of the model, design factors, and possible violations of the assumptions of SEM. Also, design considerations should impact the validity and the generalizability of the findings.

To facilitate the verification of the proposed theory, interpretive discussion should include study

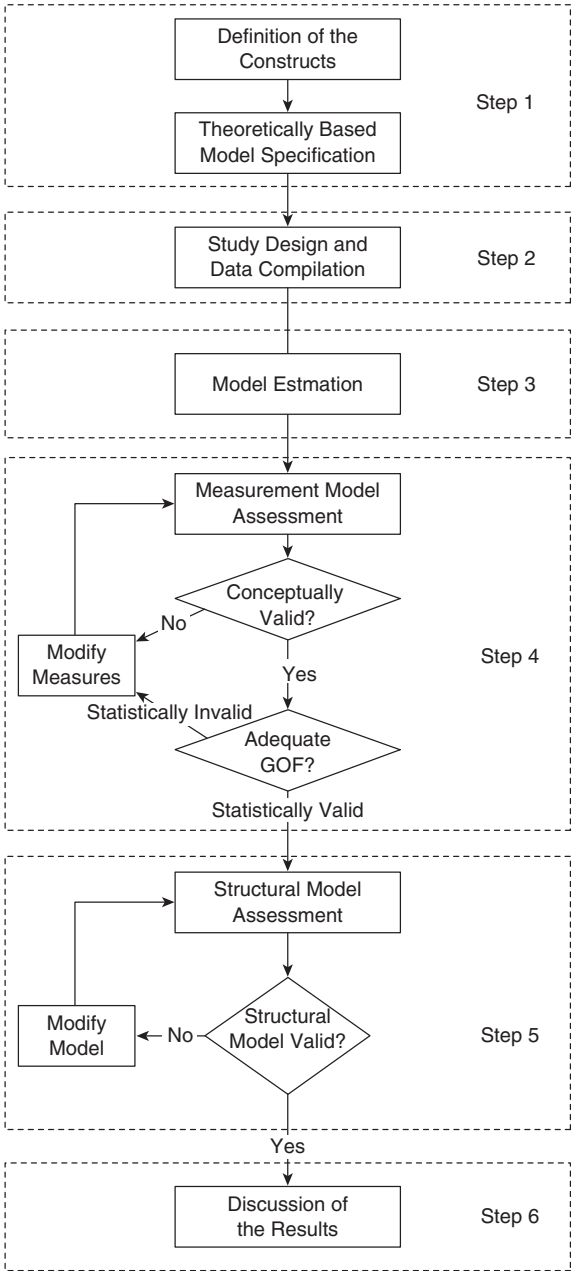



**Figure 5** SEM Process

setting, samples, measurements, or estimation decisions that might have any underlying effects on drawing the conclusions from the SEM analysis.

## Developments

The reader should be aware of the existence of methodological advances that can extend the basic

statistical qualities and relax some of the fundamental assumptions of SEM. For instance, nonnormality can be addressed using WLS estimators. Besides, current developments in SEM allow nonlinearity of the observed variables and existence of interactions among latent variables. Two-stage least square (2SLS) estimation is a solid alternative to MLE to account for interactions. It is also useful when the model contains feedback loop(s). A longitudinal analysis of data can be performed in SEM using the latent growth curve model (LGM).

*Ruben Mancha and Mark T. Leung*

***See also*** Cause and Effect; Confirmatory Factor Analysis; Construct Validity; Experimental Design; Multivariate Normal Distribution; Significance Level, Interpretation and Construction; Quasi-Experimental Design

**Further Readings**


Browne, M. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical & Statistical Psychology, 37,* 62–83.

Hair, J., Black, W., Babin, B., Anderson, R., & Tatham, R. (2006). *Multivariate data analysis* (6th ed.). Upper Saddle River, NJ: Prentice Hall.

Jöreskog, K. G. (1977). Structural equations models in the social sciences: Specifications, estimation and testing. In P. R. Krishnaiah (Ed.), *Applications of statistics* (pp. 265–287). Amsterdam, the Netherlands: Holland.

Jöreskog, K. G. (1981). Basic issues in the application of LISREL. *Data, 1,* 1–6.

Jöreskog, K. G., & Sorbom, D. (1993). LISREL III: *Estimation of linear structural equation systems by maximum likelihood method.* Chicago: National Educational Resources.

Kaplan, D. (2000). *Structural equation modeling: Foundations and extensions.* Thousand Oaks, CA: Sage.

Muthen, B. (1983). Latent variable structural equation modeling with categorical data. *Journal of Econometrics, 22,* 48–65.

Muthen, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49,* 115–132.

Schumacker, R., & Lomax, R. (1996). *A beginner's guide to structural equation modeling.* Mahwah, NJ: Lawrence Erlbaum.

Schumacker, R., & Marcoulides, G. (1998). *Interaction and nonlinear effects in structural equation modeling.* Mahwah, NJ: Lawrence Erlbaum.

Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Alexandria, VA: American Statistical Association.

Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology, 15,* 201–293.

Wright, S. (1918). On the nature of size factors. *Genetics, 3,* 367–374.


# Student's *t* Test

The Student's *t* test is, arguably, the most used statistical procedure. Because it is, by far, the most frequently used test for comparing differences between sample means for two independent groups (e.g., a treatment group receiving a treatment vs. a control group receiving no treatment), or when comparing average performance over time (e.g., before treatment and after treatment), this entry first introduces the test in these two contexts.

## Inferences About $\mu_1 - \mu_2$ Based on Independent Samples

The null and alternative hypotheses for a two-sided test (i.e., a nondirectional test) regarding equality of population means $\mu_j$s ($j = 1,2$) for the two groups (e.g., treatment [group 1] versus control [group 2]) are:

$$H_0 : \mu_1 - \mu_2 = k,$$
$$H_1 : \mu_1 - \mu_2 \neq k.$$

The directional hypotheses are:

$$H_0 : \mu_1 \geq \mu_2 = k,$$
$$H_1 : \mu_1 < \mu_2 \neq k,$$

or

$$H_0 : \mu_1 \leq \mu_2 = k,$$
$$H_1 : \mu_1 > \mu_2 \neq k.$$

To test either null hypothesis (only one set of null and alternatives hypotheses are used in any one situation), the Student's two independent sample test statistic is the statistic of choice, and is equal to

$$t^* = \frac{(\overline{X}_1 - \overline{X}_2) - \mathrm{E}(\overline{X}_1 - \overline{X}_2)}{\sqrt{\text{est } \sigma^2_{\overline{X}_1 - \overline{X}_2}}},$$

where

- $\overline{X}_1$ and $\overline{X}_2$ are the means of the samples of observations from populations one and two, respectively,
- $E(\overline{X}_1 - \overline{X}_2)$ refers to the expected value of the difference between the two sample means, a value given by the null hypothesis, which typically is stipulated to be zero (though the null hypothesis can be that the population difference equals any real number, e.g., $H_0 : \mu_1 - \mu_2 = 100$), and
- est $\sigma^2_{\overline{X}_1 - \overline{X}_2}$ stands for the estimate of the variance of the difference between two independent sample means.

Because the standard deviation of the difference between two independent sample means equals

$$\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}, \sqrt{\text{est } \sigma^2_{\overline{X}_1 - \overline{X}_2}} = \sqrt{\text{est}\left(\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}\right)}.$$

The test statistic was derived under the restriction that the population variances for the groups are equal. Accordingly, the statistic can be expressed as

$$t^* = \frac{(\overline{X}_1 - \overline{X}_2) - E(\overline{X}_1 - \overline{X}_2)}{\sqrt{\text{est } \sigma_0^2\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}},$$

where $\sigma_0^2$ is the common value of variance, and $N_1$ and $N_2$ are the sample sizes for groups 1 and 2, respectively.

Thus, the *t* statistic, given the usual null hypothesis of equal population means, is equal to

$$t^* = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\text{est } \sigma_0^2\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}},$$

where

$$\text{est } \sigma_0^2 = s_p^2 = \frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}$$

and $s_1^2$ and $s_2^2$ are the unbiased estimates of the common population variance from groups 1 and 2, respectively, and are thus pooled into one estimate, $s_p^2$.

William Gosset, publishing under the pseudonym of "Student," derived the probability density function for a *t* variable, which is

$$f(t) = G(\nu)\left[1 + \frac{t^2}{\nu}\right]^{-(\nu+1)/2}.$$

This probability density function is often called the *t* distribution. An examination of the density function reveals a lot about characteristics of the distribution of *t* variables. First, the probability of observing any value of *t* (the statistic), by chance and chance alone, is determined by just one parameter $\nu$ (nu). Thus, the *t* distribution is referred to as a one parameter family of distributions; that is, by knowing $\nu$, which is also referred to as the degrees of freedom, one can determine the probability of having observed a value of *t*, that is, the value of the test statistic. One should also note that the value of *t* enters the density function as a squared value; consequently, the distribution of *t* values must be symmetric; that is, positive and negative values of *t* with the same squared value have the same probability of occurrence. In addition, because all the constants in the density function are positive and the term involving *t* (in the square brackets) is raised to a negative value, the largest positive density value is when $t = 0$. From the previous equation, it can be observed that that the *t* distribution is a symmetric unimodal distribution, bell shaped in form, resembling a normal distribution.

Student's two independent sample *t* statistic is distributed as a *t* variable with degrees of freedom ($\nu$) equal to $N_1 + N_2 - 2$. Degrees of freedom have been defined as either the number of observations that are free to take on any numerical value or the number of independent observations minus the number of parameters estimated from the data. In the context of Student's two independent sample *t* test, using our first definition of degrees of freedom, there are $N_1 - 1$ free observations when computing the variance for the scores in group 1

and $N_2 - 1$ for group 2. (This is the case because the sum of deviation scores around the mean, the numerator in computing a variance, sums to zero.) Accordingly, $(N_1 - 1) + (N_2 - 1) = N_1 + N_2 - 2$.

Let $t_{(\theta, \nu)}$ denote the point in the $t$ distribution such that the probability of a $t$ value below that point is $\theta$. To assess statistical significance one uses one of the following decision rules.

Fail to reject $H_0 : \mu_1 - \mu_2 = 0$ if $|t^*| < t_{(1-\alpha/2, \nu)}$;

Reject $H_0 : \mu_1 - \mu_2 \geq 0$ and conclude $\mu_1 - \mu_2 \neq 0$ if $|t^*| \geq t_{(1-\alpha/2, \nu)}$;

or

Fail to reject $H_0 : \mu_1 - \mu_2 = 0$ if $t^* > t_{(\alpha, \nu)}$;

Reject $H_0 : \mu_1 - \mu_2 \leq 0$ and conclude $H_1 : \mu_1 - \mu_2 < 0$ if $t^* \leq t_{(\alpha, \nu)}$;

or

Fail to reject $H_0 : \mu_1 - \mu_2 = 0$ if $t^* < t_{(1-\alpha, \nu)}$;

Reject $H_0 : \mu_1 - \mu_2 = 0$ and conclude $H_1 : \mu_1 - \mu_2 > 0$ if $t^* \geq t_{(1-\alpha, \nu)}$.

In each of these expressions, $\alpha$ is the probability of a Type I error, that is, the probability of rejecting $H_0$ when it is correct. In principle, a researcher should choose $\alpha$ before collecting the data by considering the maximum probability he or she is willing to risk for falsely rejecting $H_0$. In practice $\alpha$ is usually set at .05. Choosing $\alpha$ sets the critical value $\left[t_{(1-\alpha/2, \nu)}, t_{(\alpha, \nu)}, \text{ or } t_{(1-\alpha, \nu)}\right]$ in the Student's $t$ distribution for deciding if the data support $H_1$.

Gosset derived this statistic and density function under the conditions that the data in the population are normal (the normality assumption) in form (which results in the random variables in the numerator and denominator of the statistic being independent of one another). As well, in this two-group case, there is the requirement that the population variances ($\sigma_1^2 = \sigma_2^2 = \sigma_0^2$) are equal to one another and thus equal to a constant value (the homogeneity of variance assumption), and that the data points are independent of one another (the independence of observations assumption). Collectively, these conditions are often referred to as the validity requirements of the test statistic. The importance of the validity assumptions is that they must be satisfied for the probabilities associated with any particular value of $t$ to be accurate (i.e., exact). When one or more assumptions do not hold, then the probabilities are not exact, and accordingly, the probability of a value of $t$ is equal to some other value than what would be obtained from the density function.

The issue of whether derivational assumptions of a statistic are satisfied, or not, for some set (population) of observations, has been a matter of concern for decades. As noted earlier, the researcher selects $\alpha$ before collecting the data, and $\alpha$ determines the critical value $\left[t_{(1-\alpha/2, \nu)}, t_{(\alpha, \nu)}, \text{ or } t_{(1-\alpha, \nu)}\right]$ in the Student's $t$ distribution. If, for example, the critical value is $t_{(1-\alpha/2, \nu)}$ (see equation presented previously) and the derivational assumptions are correct, then the probability that $|t^*|$ exceeds $t_{(1-\alpha/2, \nu)}$ is guaranteed to be $\alpha$. But if the derivational assumptions are not correct, the critical value might be incorrect. Clearly, then, if a certain critical value in the distribution of $t$ (e.g., one that cuts off the lower and upper 2.5% of the distribution) [i.e., $1 - (.05/2)$] is not the correct (exact) value, because one or more of the validity assumptions did not hold, the critical values (numerical values) of $t$ that cut off the upper and lower 2.5% of the distribution are inaccurate. Correspondingly, the probability of falsely rejecting the null hypothesis is unknown—meaning that the actual probability of committing a Type I error can be less than or greater than one wants. This situation is undesirable because an experimenter wants to control this type of mistake associated with hypothesis testing. A later section of this entry addresses alternative courses of action that experimenters can adopt to circumvent the biasing effects of non-normality and/or variance heterogeneity.

### *Confidence Intervals and Power (Sensitivity) to Detect Treatment Effects*

Two additional issues should be discussed, briefly, within the context of computing any test statistic: setting an interval around the unknown

population difference and determining the required sample size to detect treatment effects.

With respect to the first issue, it should be noted that many authorities on the analysis of data strongly recommend that in addition to testing hypotheses, researchers should also compute confidence intervals (CIs) around the unknown population parameters being examined. (The zeitgeist certainly in the social sciences is that CIs and effect size statistics are paramount to understanding the phenomenon under investigation.) The CI takes into account sampling variability in that it specifies a range of values that are equally reasonable as estimates of the unknown parameter. (In this case the difference between two means.)

For the case with two independent groups, a $100(1-\alpha)\%$ CI on $\mu_1 - \mu_2$ around $\overline{X}_1 - \overline{X}_2$ is formed as follows:

$$(\overline{X}_1 - \overline{X}_2) \pm t_{[1-(\alpha/2),\nu]} s_{\overline{X}_1 - \overline{X}_2}.$$

It is also interesting to note that values that fall within the interval would not have been rejected if they had been specified as $k$ in the null hypothesis $H_0 : \mu_1 - \mu_2 = k$. One might also compute one-sided CIs:

$$(\overline{X}_1 - \overline{X}_2) + t_{[1-\alpha,\nu]} s_{\overline{X}_1 - \overline{X}_2}$$
$$(\overline{X}_1 - \overline{X}_2) - t_{[\alpha,\nu]} s_{\overline{X}_1 - \overline{X}_2}$$

where $s$ stands for standard deviation.

The second issue that warrants some discussion is the power to detect treatment effects when they are present. In addition to attending to limiting Type I errors, researchers should, when designing a study, give considerable attention to what they must do to optimize their chances of detecting a treatment effect that might be present.

Several factors affect the sensitivity of a statistic to detect treatment effects when present, and they include the following: (a) the magnitude of the true effect, (b) the level of significance ($\alpha$), (c) the amount of noise (error) in the system (all extraneous noncontrolled sources of variation), and (d) sample size. Factors (a) and (c) should be optimized in any research endeavor if researchers show due diligence with regard to controlling all known sources of extraneous variance and creating independent variables that potentially maximize the effect that they are investigating. Even though manipulating the level of significance is most reasonable for increasing the power to detect real treatment effects (e.g., increasing the value of $\alpha$, increases the power to detect effects), researchers tend to be reluctant to do so. Thus, the one factor that is most amenable for increasing the power of a statistical test is the size of sample. Increasing sample size increases the power to detect effects (all other factors constant). In the past, most treatises on determining sample size to increase the power to detect treatment effects discussed formula and tables or charts to determine how many subjects were required in the study (that is, $N = N_1 + N_2$); this approach involved specifying the magnitude of the treatment that one desires to detect, that is, specifying the noncentrality parameter. To examine treatment effects, a different probability distribution of $t$ is required—the noncentral $t$ distribution; the previously presented $t$ distribution, which exists under the null hypothesis, is referred to as the central $t$ distribution. However, many in the field now rely on statistical software to determine the required sample size. An excellent program, G*POWER, has been written by Franz Faul and Edgar Erdfelder, and it can be obtained from the G*POWER website.

## Inferences About $\mu_1 - \mu_2$ Based on Dependent (Correlated) Samples

The inferential procedure (i.e., comparing the Student's $t$ statistic to the critical value) for assessing whether the means in the population are equal for two dependent groups (e.g., comparing husbands and wives on an attitudinal measure) or equal across two times of measurement (e.g., assessing the number of trials to complete a task and subsequently reassessing the subjects after a 2-week interval) is similar to the case with two independent groups. The only difference is the formula for Student's test statistic and the degrees of freedom. In the dependent groups scenario, the data analysis and the test statistic can be conceptualized from either a one-sample or two-sample approach.

### *One-Sample Conceptualization for the Analysis*

For this approach the data for, say, husbands $(X_{i1} : i, \ldots, N)$ and wives $(X_{i2} : i, \ldots, N)$ is used to

create a new variable $d_i = X_{i1} - X_{i2}$, which is a difference score; thus, a set of scores from two samples has now become a set of scores from one sample. A test statistic is constructed in the usual manner, that is,

$$t^* = \frac{\overline{d} - E(\overline{d})}{\text{est } \sigma_{\overline{d}}},$$

where

$\overline{d}$ is the sample mean of the $d_i$s,

$E(\overline{d})$ is the expected value of $\overline{d}$, the value specified under the null hypothesis (which is typically zero), and

est $\sigma_{\overline{d}}$ is the estimate of the standard deviation of $\overline{d}$.

After substituting a sample estimate for the unknown standard deviation and assuming the usual null hypothesis that the mean of the difference scores is equal to zero in the population, the test statistic is defined as

$$t^* = \frac{\overline{d}}{s_{\overline{d}}},$$

where $s_{\overline{d}} = \frac{s_d}{\sqrt{N}}$ and $s_d$ is the unbiased variance of the difference scores. This $t$ statistic is distributed as a $t$ variable with degrees of freedom ($\nu$) equal to $N - 1$, where $N$ equals the number of difference scores.

Accordingly, the decision rules relating to the null and alternative hypotheses are (at this point, and on, to save space, only nondirectional hypotheses are enumerated.)

Fail to reject $H_0 : \mu_{d_i} = 0$ if $|t^*| < t_{(1-\alpha/2,\,\nu)}$.

Reject $H_0 : \mu_{d_i} = 0$ and conclude $\mu_{d_i} \neq 0$ if $|t^*| \geq t_{(1-\alpha/2,\,\nu)}$.

### *Two-Sample Conceptualization for the Analysis*

When conceptualized as a two-group problem, the standard deviation of $\overline{X}_1 - \overline{X}_2$ for dependent samples is different than in the two independent group case because the correlation between the two group of scores (husbands vs. wives; measurements at time one vs. those at time two) must be taken into account. It can be shown that this standard deviation is equal to

$$\sigma_{\overline{X}_1 - \overline{X}_2} = \sqrt{\sigma^2_{\overline{X}_1} + \sigma^2_{\overline{X}_2} - 2\rho\sigma^2_{\overline{X}_1}\sigma^2_{\overline{X}_2}},$$

where

$\sigma^2_{\overline{X}_1}$ and $\sigma^2_{\overline{X}_2}$ are the population variances of the means for groups 1 and 2, respectively, and

$\rho$ is the correlation between the two sets of measurements. (In fact, $\rho\sigma^2_{\overline{X}_1}\sigma^2_{\overline{X}_2}$ equals the covariance between $\overline{X}_1$ and $\overline{X}_2$.)

The test statistic can be represented as

$$t^* = \frac{(\overline{X}_1 - \overline{X}_2) - E(\overline{X}_1 - \overline{X}_2)}{\sqrt{\text{est}\left(\sigma^2_{\overline{X}_1} + \sigma^2_{\overline{X}_2} - 2\rho\sigma^2_{\overline{X}_1}\sigma^2_{\overline{X}_2}\right)}}.$$

After replacing the unknown parameters with their least squares sample estimates and assuming the usual null hypothesis of a zero difference between the dependent means, the statistic equals

$$t^* = \frac{(\overline{X}_1 - \overline{X}_2)}{\sqrt{s^2_{\overline{X}_1} + s^2_{\overline{X}_2} - 2rs^2_{\overline{X}_1}s^2_{\overline{X}_2}}},$$

where $s^2_{\overline{X}_1}$ and $s^2_{\overline{X}_2}$ equal $\frac{s^2_{X_{i1}}}{N}$ and $\frac{s^2_{X_{i2}}}{N}$, respectively, and $r$ is the sample estimate of the population correlation coefficient.

As should be the case,

$$\frac{s_d}{\sqrt{N}} = \sqrt{s^2_{\overline{X}_1} + s^2_{\overline{X}_2} - 2rs^2_{\overline{X}_1}s^2_{\overline{X}_2}}.$$

The decision rule to assess statistical significance is

Fail to reject $H_0 : \mu_1 - \mu_2 = 0$ if $|t^*| < t_{(1-\alpha/2,\,\nu)}$.

Reject $H_0 : \mu_1 - \mu_2 = 0$ and conclude $\mu_1 - \mu_2 \neq; 0$ if $|t^*| \geq t_{(1-\alpha/2,\,\nu)}$,

where $\nu = N - 1$, and $N$ is the number of paired observations.

### *Confidence Intervals and Power (Sensitivity) to Detect Treatment Effects*

As was the case in the two independent groups scenario, CIs and power analyses should also be considered by experimenters when comparing two dependent groups. The two-sided CI for $\mu_{d_i}$ is

$$\overline{d} \pm t_{(1-\alpha/2,\nu)} s_{\overline{d}},$$

while the one-sided intervals are

$$\overline{d} + t_{(1-\alpha/2,\nu)} s_{\overline{d}}$$
$$\overline{d} - t_{(\alpha,\nu)} s_{\overline{d}}.$$

The intervals from the two-sample perspective merely replace $(\overline{X}_1 - \overline{X}_2)$ for $\overline{d}$ and $s_{(\overline{X}_1-\overline{X}_2)}$ for $s_{\overline{d}}$. G*POWER can once again be used to determine the required number of observations (pairs) that are required to detect an effect of a given size.

## Additional Applications

In addition to being used to test hypotheses about pairs of means, *t* test statistics and the Student's *t* distribution are used in many other situations. For example, when the experimenter needs to test hypotheses about three or more means, a need that arises in designs with more than two treatments, it is common for experimenters to test contrasts, where a contrast is a weighted sum of means in which the weights sum to zero. For example, if an experiment involved two active treatments (with population means $\mu_1$ and $\mu_2$) and a control treatment (with population mean $\mu_c$), then the researcher might be interested in testing each of the possible pairwise comparisons (e.g., $H_0 : \mu_1 - \mu_c = 0$) or a complex comparison such as $H_0 : .5\mu_1 + .5\mu_2 - \mu_c$, which compares the average of the means for the two active treatments with the mean for the control treatment. Each of these hypotheses can be tested by using a *t* test statistic. The choice of critical value is a complex issue, but in many cases will be a point of Student's *t* distribution.

The Student's test statistic and distribution are also used in testing hypotheses about Pearson product-moment correlation coefficients, partial correlation coefficients, and regression coefficients. For example, if $\rho$ is the population Pearson product-moment correlation coefficient and the researcher is interested in testing $H_0 : \rho = 0$, then the Student's test statistic is

$$t^* = r\sqrt{\frac{N-2}{1-r^2}},$$

where $r$ is the sample Pearson product-moment correlation and $N$ is the number of pairs of scores. The rules for determining whether $t$ is significant are the same as when means are compared; however, the degrees of freedom are $\nu = N - 2$. As these few examples illustrate, the *t* test statistics and Student's *t* distribution are widely used in research.

## Robust Alternatives to the Classic *t* Statistic

As previously indicated, the two independent sample *t* statistic was derived assuming that the data in the population are normally distributed, that the variances of the two populations are equal (and equal to one constant value), and that the observations are independent of one another. The last requirement is typically not an issue when experimenters follow good experimental procedures (e.g., not allowing the result from any subject to be influenced by that of another subject). However, with regard to normality and homogeneity of variances, most commentators of validity requirements believe that data rarely, if ever, conform to these validity requirements. Thus, as mentioned previously, the probability of a Type I error will not equal the theoretical value (i.e., $\alpha \neq .05$).

For decades, many authors have proposed solutions that are intended to obviate the deleterious effects of derivational assumptions not being satisfied. This issue is generically referred to as the robustness of a statistical test. Robust statistical tests are those that are generally insensitive to assumption violations. In other words, the actual probability of a Type I error and the nominal value (the value set by the experimenter) are similar in value, where nonrobust tests are adversely affected by assumption violations (i.e., there is a discrepancy between the actual and nominal values of significance).

With regard to variance heterogeneity, remember that in the case with two independent groups, the *t* statistic can be defined as

$$t^* = \frac{(\overline{X}_1 - \overline{X}_2) - E(\overline{X}_1 - \overline{X}_2)}{\sqrt{\text{est}\left(\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}\right)}}.$$

We simplified this statistic by assuming equality of variances; factored out, accordingly, the common variance; and used a pooled value to estimate this unknown common population variance. However, when variances cannot be presumed to be equal, it has been recommended that group variances not be pooled and one merely estimates $\sigma_1^2/N_1$ and $\sigma_2^2/N_2$. Testing the usual null hypothesis, and replacing unknown population parameters with sample estimates, the test statistic equals

$$t_W^* = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}},$$

which is approximately distributed as a $t$ variable with degrees of freedom equal to

$$\nu_W = \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{\left(\frac{s_1^2}{N_1}\right)^2}{N_1 - 1} + \frac{\left(\frac{s_2^2}{N_2}\right)^2}{N_2 - 1}},$$

where $W$ in the subscripting of the statistic and degrees of freedom refers to Welch, an author who provided one solution (perhaps the most widely used one) for comparing two independent groups in the possible presence of unequal variances. The decision rule for assessing the null and alternative hypotheses is

$$\text{Fail to reject } H_0 : \mu_1 - \mu_2 = 0 \text{ if } |t^*| < t_{(1-\alpha/2,\, \nu_W)}.$$

$$\text{Reject } H_0 : \mu_1 - \mu_2 = 0 \text{ and conclude } H_1 : \mu_1 - \mu_2 \neq 0 \text{ if } |t^*| \geq t_{(1-\alpha/2,\, \nu_W)}.$$

One-sided tests can also be examined.

The empirical literature indicates that the Welch solution typically is not substantially affected by heterogeneous variances (i.e., the actual and nominal values for the Type I error rate are not too discrepant from one another), even when group sizes are unequal. Moreover, the power of this statistic to detect non-null treatment effects compares favorably with the regular $t$ statistic even when variances happen to be equal, a case in which the regular test should be uniformly the most powerful.

The reader should note that Welch's solution is for testing mean equality in the presence of unequal variances and does not address the issue of non-normality. That is, not only might the population variances be unequal, but the observations might as well be non-normal in shape. The empirical literature indicates that when data are non-normal, the actual probability of committing a Type I error might deviate from the nominal value (i.e., an inexact test), and, as well, the power to detect effects might be substantially diminished. For this scenario, many solutions have been proposed since the 1950s. One solution, namely replacing the usual least squares estimators for the mean and variance with robust estimators—trimmed means and Winsorized variances, and applying these estimators with Welch's nonpooled statistic, is discussed.

To compute a trimmed mean, one removes an a priori determined percentage of the largest and smallest observations and computes the mean from the remaining observations. The population trimmed mean is denoted by $\mu_{tj}$ and the corresponding sample estimate as $\overline{X}_{tj}$. A common trimming percentage is 20%, which means that, in total, 40% of the data are trimmed (for a justification of 20% trimming, see Wilcox (2005, p. 57).

When trimmed means are introduced to researchers, a common reaction is that one throws away data when the trimmed means are used, and therefore the trimmed mean cannot be a better choice than the usual least squares mean. But the trimmed mean has advantages that the least squares mean does not. Consider a situation in which the sampled distribution is symmetric. Then, both the trimmed mean and least squares mean estimate the same parameter, the population mean, and an important criterion for selecting between the two estimators is the standard error. It can be shown that when the data are sampled from a long-tailed distribution, the trimmed mean will have a smaller standard error than does the mean. Granted, if the data are sampled from a short-tailed or normal distribution, the mean will have a smaller standard error. However, the advantage for the mean

under these conditions is often not large, whereas the advantage for the trimmed mean can be substantial when the data are sampled from a long-tailed distribution. Thus, one argument for the trimmed mean is that it can have a substantial advantage in terms of accuracy of estimation when sampling from long-tailed symmetric distributions, although it has a slight disadvantage when sampling from normal and short-tailed distributions.

Additionally, the trimmed mean is preferable when the data are subject to outliers. In contrast, the mean is not resistant to outliers, and a single outlying observation can destroy its effectiveness as an estimator. If the distribution from which the data are sampled is skewed, then the argument against the mean is that it is not a good indicator of the central point of the distribution because the extreme scores drag it away from the center. Because the trimmed distribution will be more nearly symmetric than the nontrimmed distribution, the trimmed mean will be a better indicator of the center of the distribution. Some researchers find these arguments for using the trimmed mean, rather than the mean, persuasive; however, others might not. But researchers who use the mean should be aware that they are using an estimator that has poor accuracy (in some situations), is not resistant to outliers, and might not be a good indicator of the center of the distribution.

To compute a Winsorized variance, the smallest nontrimmed score replaces the scores trimmed from the lower tail of the distribution, and the largest nontrimmed score replaces the scores removed from the upper tail. The nontrimmed and replaced scores are called Winsorized scores. A Winsorized mean is calculated by applying the usual formula for the mean to the Winsorized scores, and a Winsorized variance is calculated as the sum of squared deviations of Winsorized scores from the Winsorized mean divided by $N_j - 1$. The Winsorized variance is used because it can be shown that the standard error (deviation) of a trimmed mean is a function of the Winsorized variance. A sample Winsorized variance is denoted by $s^2_{W_j}$. See Wilcox (2005) for formulas corresponding to the verbal descriptions of the trimmed mean and Winsorized variance.

Let

$$\tilde{s}^2_{W_j} = \frac{(n_j - 1)s^2_{W_j}}{h_j - 1}$$

be the scaled Winsorized variance, where $h_j$ stands for the effective sample size, that is, the size after trimming the data. In computing the robust Welch–James (WJ) statistic with robust estimators, the quantity $\tilde{s}^2_{W_j}$ plays the role that $s^2_j$ plays in computing the WJ statistic based on least squares estimators. Specifically, the robust WJ statistic is

$$t^*_{WJt} = \frac{\overline{X}_{1t} - \overline{X}_{2t}}{\sqrt{\frac{\tilde{s}^2_{W_1}}{h_1} + \frac{\tilde{s}^2_{W_2}}{h_2}}}.$$

The critical value is obtained from the $t$ distribution by using the estimated $df$

$$\hat{\nu}_t = \frac{\left(\frac{\tilde{s}^2_{W_1}}{h_1} + \frac{\tilde{s}^2_{W_2}}{h_2}\right)^2}{\frac{\tilde{s}^4_{W_1}}{h_1^2(h_1 - 1)} + \frac{\tilde{s}^4_{W_2}}{h_2^2(h_2 - 1)}}.$$

Thus far, this entry has illustrated inferential procedures that differ in terms of the means and variances used to compute the $t$ statistic but that use the theoretical $t$ distribution to obtain the critical value; however, a nonparametric bootstrap can be used to obtain the critical value. The strategy behind the use of the bootstrap in hypothesis testing is to shift the sample distributions of the scores for each group and variable by subtracting the group mean (least squares or trimmed) from each score and using the shifted empirical distributions to estimate an appropriate critical value.

In particular, for each $j$, obtain a bootstrap sample by randomly sampling with replacement $N_j$ observations from the shifted values, yielding $X^*_1, \ldots, X^*_{N_j}$. Let $t^*$ be the value of a test statistic ($t_W$) or ($t_{WJ_t}$) based on the bootstrap sample. For a two-tailed test, the $B$ values of $|t^*|$, where $B$ represents the number of bootstrap simulations, are put in ascending order, that is, $\left|t^*_{(1)}\right| \leq \cdots \leq \left|t^*_{(B)}\right|$, and an estimate of an appropriate critical value is $\left|t^*_{WJ_{t(q)}}\right|$, $q = (1 - \alpha)B$, rounded

to the nearest integer. For example, if $\alpha = .05$ and the number of bootstrap samples is 1,000, then $q = (1 - \alpha)B$ is 950 and $\left|t^*_{WJt(q)}\right|$ is the 950th value in the rank order of the $B$ values of $|t^*|$. One will reject $H_0 : \mu_1 - \mu_2 = 0$ or $H_0 : \mu_{1t} - \mu_{2t} = 0$ when $|t^*| \geq \left|t^*_{(q)}\right|$, where $|t^*|$ is the value of the heteroscedastic statistic ($t^*_W$ or $t^*_{WJt}$) based on the original nonbootstrapped data.

CIs and power analyses can also be performed with this robust $t$ test statistic (see Wilcox, 2005; Luh & Guo, 2007).

*H. J. Keselman and James Algina*

***See also*** Significance Level, Concept of; *t* Test, Independent Samples; *t* Test, One Sample; *t* Test, Paired Samples

### Further Readings

James, G. S. (1951). The comparison of several groups of observations when the ratios of the population variances are unknown. *Biometrika, 38,* 324–329.

Johansen, S. (1980). The Welch-James approximation to the distribution of the residual sum of squares in a weighted linear regression. *Biometrika, 67,* 85–92.

Lix, L. M., & Keselman, H. J. (1995). Approximate degrees of freedom tests: A unified perspective on testing for mean equality. *Psychological Bulletin, 117,* 547–560.

Luh, W., & Guo, J. (2007). Approximate sample size formulas for the two-sample trimmed mean test with unequal variances. *British Journal of Mathematical and Statistical Psychology, 60,* 137–146.

Student. (1908). The probable error of a mean. *Biometrika, 6,* 1–25.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika, 29,* 350–362.

Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing* (2nd ed.). New York: Elsevier.

Wilcox, R. R., & Keselman, H. J. (2003). Modern robust data analysis methods: Measures of central tendency. *Psychological Methods, 8,* 254–274.

### Websites

G*POWER: http://www.psychologie.uni-duesseldorf.de/aap/projects/gpower

## Sums of Squares

In mathematics, the sum of squares can refer to any set of numbers that has been squared and then added together (i.e., $\sum X^2$). However, in the statistical analysis of quantitative data, the sum of squared differences or deviations (normally the difference between a score and the mean) is of particular interest. This formulation is usually referred to in research by the term *sums of squares*. This type of sum of squared values is extremely important in the analysis of the variability in data and in understanding the relationship between variables.

If there is no variability in a set of numbers, then they are all the same. However, this is unlikely to be the case in any research. Indeed, researchers actively investigate changes in their data resulting from experimental manipulations. So a procedure is required to produce a value to express the amount of variability within a data set. The sum of squares calculation achieves this outcome. For example, consider the numbers 2, 4, and 6. The range gives us a crude measure of the spread of these numbers: between 2 and 6 we have a range of 4, but this ignores the distribution of the scores within that range. More subtly, to indicate the variation within the data, we can compare each score to the mean ($\overline{X}$), which in this case is 4, to produce a difference or deviation from the mean. So the deviations are $2 - 4 = -2$, $4 - 4 = 0$, and $6 - 4 = 2$. It could be assumed that simply adding up these deviations would provide a useful measure of total variation in the data set. Unfortunately, this is not the case, as this total will always be zero, with the deviations of numbers below the mean canceling out the deviations of the numbers above the mean, as in the previous example, where $-2 + 0 + 2 = 0$. A solution to this problem is to square the deviations before they are summed, which always produces a positive nonzero value when there is some variability in the data. For the example the calculation is $(-2)^2 + 0^2 + 2^2 = 8$. This is now a measure of the variability within the data, which can be expressed mathematically as $\sum (X - \overline{X})^2$, the sum of the squared deviations from the mean, or the sum of squares. This is an extraordinarily useful measure and is integral to most of the statistical techniques used to analyze quantitative data.

Sums of squares are absolutely critical to quantitative data analysis, as variability in the data can be calculated and attributed to different sources, which allows researchers to make appropriate judgments about the relationship between the variables under investigation.

## Variance and Standard Deviation

The previous sum of squares formula lies at the heart of the most common statistic for expressing variability within research data, the standard deviation. With a population of numbers, the mean squared deviation can be calculated by dividing the sum of squares by the number of scores $n$. This mean square value is called the variance. However, in research, samples are normally selected to estimate population parameters so, more usually, the sum of squares is divided by the degrees of freedom $(n-1)$ to produce the variance,

$$\frac{\sum (X-\overline{X})^2}{n-1},$$

as this provides a better estimation of the population value. The square root of the variance is referred to as the standard deviation:

$$\sqrt{\frac{\sum (X-\overline{X})^2}{n-1}}, \text{ which is } \sqrt{\frac{\text{sum of squares}}{\text{degrees of freedom}}}.$$

This is the standard measure of the variation of the scores in a sample. In the previous example, the standard deviation is 2. (Note that the sum of squares is often expressed by

$$\sum X^2 - \frac{(\sum X)^2}{n}$$

in the formula for the variance or standard deviation. This is equivalent to

$$\sum (X-\overline{X})^2.)$$

## Linear Correlation and Regression

Sums of squares are also central in the calculation of linear correlation and regression, as a measure of the variability in the data. Note that in a correlational analysis, researchers are looking to determine whether there is a relationship between two or more variables—whether they vary together—by examining the variation within the research data. For example, it might be predicted, not surprisingly, that the amount of time spent studying correlates positively with performance in an examination. On plotting a graph (a scatterplot) of "study time" versus "examination result" for a group of students, it is observed that, generally, the more a student studies, the higher the result in the examination, indicating a positive correlation. However, the results are unlikely to all lie neatly along a straight line on the graph. More usually, the scores fall within a broad band. Yet the prediction is that the relationship between the two variables is linear (i.e., predicting a linear model), and so it is argued that the reason any result deviates from the straight line is *error* (i.e., random variation). To calculate the best linear model for the data (i.e., the regression line), it is necessary to find the line that produces the smallest overall error. This is done by using the sum of squares formula. With study time as the $X$ axis and examination result as the $Y$ axis, the sum of squares formula becomes $\sum (Y-Y')^2$, where $Y$ is the actual examination result and $Y'$ is the predicted examination result lying on the straight line, the linear model. The sum of squares now gives a measure of the overall deviation of the points from that predicted straight line. To find the line of "best fit" for these data, a line that minimizes this sum of squares must be computed. To do this, a procedure called the least squares method is employed, and the regression line can be calculated.

The sum of squares is crucial both to the calculation of the degree of correlation (correlation coefficient, $r$), and in the calculation of the regression line. Within the formulas for each of these calculations lie measures of variability in the data: sums of squares. A correlation coefficient of $+1$ indicates that there is no error or residual data to be accounted for, as all the points on the graph lie along a straight line. However, we are unlikely to find this with research data. Even a strong correlation will still have error. Here, error, or residual, indicates the portion of each result that cannot be explained by the linear model (the regression line). Imagine one student studied for 20 hours and scored 63 in the examination. From the calculation of the regression line, it is predicted that anyone

studying for 20 hours will receive a score of 61 in the examination. Thus, for this student, 61 of the 63 score can be explained by the regression, but the residual or "extra" variability in the score of 2 cannot be explained by the model so has to be attributed to error variability. Calculating the sums of squares for the regression: $\sum (Y' - \overline{Y})^2$ (i.e., the variation in the scores explained by the regression line, where $\overline{Y}$ is the mean examination result) and dividing this by the sums of squares for the observed examination results: $\sum (Y - \overline{Y})^2$, shows how much variability in the actual examination results can be explained by the regression line. Actually this calculation is called the coefficient of determination and equals $R^2$, the square of the correlation coefficient. Thus, a correlation of $r = 0.7$ gives a coefficient of determination of $R^2 = 0.49$, indicating that 0.49% of the variation in the examination results can be explained by the regression line and 0.51% of the variation in the data is not explained by this model. The key point here is that the calculation of different sums of squares produces this information.

## The Analysis of Variance

The comparison of sums of squares in explaining the variability within data is crucial to the analysis of variance, which is one of the most important statistical techniques of quantitative data analysis. In a correlation, continuous variables are jointly examined, such as the relationship between study time and examination performance. In an analysis of variance, the effect of different categories of an independent variable (such as different teachers) on a continuous dependent variable (such as examination performance) is studied. For example, a researcher might wish to compare the effect of different teachers on the students' examination performance at the end of the semester. The students are randomly allocated to different classes and they all follow the same syllabus to control for confounding. Again, the question to be answered by the data analysis is: What is the explanation of the variability in the data? The answer to this question lies in the calculation of sums of squares.

In the analysis of variance, the sums of squares for different sources of variation are calculated, which is the key to the analysis. For example, 20 students in each class of three classes are taught by teachers Anderson, Berkowitz, and Chavez. All students take the same examination at the end of the semester. The variability of the examination results for all 60 students can be worked out by simply finding the overall mean score and calculating the sum of squares. This is the total sum of squares or total variability in the data. However, the variability of subsets of the data can be produced by calculating different sums of squares. For example, only the 20 students in Anderson's class can be selected, the class mean obtained, and the sum of squares for this class produced. As Anderson has taught all these students, any differences between the student examination results are assumed to be caused by random factors such as individual differences rather than differences in the teacher. This can be done for the other two teachers as well. These three figures provide the variation of the examination results *within* a class. The hypothesis is that there will be systematic differences between the classes on their examination results as an effect of the teacher is predicted. The mean examination result for each class has already been calculated. By calculating a sum of squares based on these means, the variability in the scores *between* the classes—and hence between the teachers—can be produced.

Now, the average amount of variability—the variance or mean square—caused by each source of variation can be compared. The variance resulting from the differences between the classes can be compared with that arising from within the classes by dividing the first by the second variance. Essentially, this comparison is examining the amount of variability arising from the systematic differences (plus any random factors) compared with the variability arising from random factors alone. Again, random variation is often termed *error.* Clearly, if variation in the examination results arising from the difference between the classes exceeds the error variation by a significant amount, then this supports the hypothesis that the students were performing differently with different teachers.

The analysis of variance, like the regression, is based on a liner model. Just as a more complex regression analysis can be performed by a multiple regression, there are also more complex analyses of variance than the simple example considered

previously. Indeed, the analysis of variance is a very general test that deals with multiple independent variables as well as multiple dependent variables, where complex interactions can be investigated. Hence, the underlying model is referred to as the *general linear model*. The key point of any modeling is to calculate how much of the variation in any observed data can be accounted for by the model and how much is error or residual. Thus, statisticians have recommended which are the most appropriate sums of squares to use in the analysis of variance for different types of research design, in particular what to do when there are different sample sizes or when there are missing cells in the data. In all cases, they are looking for the best way of representing the research data in terms of the variability in the scores that can be explained by a model and that which cannot. Once this information is available—and the different sums of squares provide it—then a researcher can decide whether the variation explained by the model is significant or whether he or she has confidence in the model as an explanation of the variability in the data.

*Perry R. Hinton*

***See also*** Analysis of Variance (ANOVA); Bivariate Regression; Coefficients of Correlation, Alienation, and Determination; Correlation; Error; General Linear Model; Pearson Product-Moment Correlation Coefficient; Scatterplot; Standard Deviation; Variance

**Further Readings**

Handy, M., & Bryman, A. (Eds). (2004). *Handbook of data analysis*. Thousand Oaks, CA: Sage.
Hinton, P. R. (2004). *Statistics explained* (2nd ed.). London: Routledge.
Howell, D. C. (2008). *Fundamental statistics for the behavioral sciences* (6th ed.). Belmont, CA: Thomson/Wadsworth.

# SURVEY

A survey is a data-collection method in which individuals answer specific questions about their behavior, attitudes, beliefs, or emotions. Surveys are frequently used by multiple disciplines, including social and behavioral sciences, political sciences, public health, and business. Surveys are commonly used in nonexperimental (correlational) research but might also be incorporated into experiments. Nonexperimental surveys can be either cross-sectional (a single data collection), repeated cross-sectional (two or more data collections with different samples), or panel surveys (multiple data collections from the same sample). However, unless an experimental design is used, survey research does not allow for the drawing of causal inferences.

## Types of Survey Questions

Surveys are composed of multiple questions assessing the constructs of interests. Typically, questions regarding demographic characteristics of participants (e.g., age, sex, race, marital status, income, or education level) are also included. If new questions are developed, they should be first evaluated in a pilot study to ensure that they are clearly worded and correctly understood, reliable, and valid. The main types of survey questions include open-ended and closed-ended questions. The latter type can be divided into dichotomous, nominal, rank-ordered, ordinal, and continuous, depending on the characteristics of the response options.

### *Open-Ended Versus Closed-Ended Questions*

Open-ended questions require participants to formulate answers in their own words. Examples of such questions are "What is the main stressor in your life?" and "What do you think should be done to control crime in inner cities?" Space is typically provided for participants to write their answers. Although open-ended questions typically elicit more complete and deeper answers than closed-ended questions, they are much more difficult to analyze quantitatively. The answers need to be first coded into categories that are either predetermined or developed through content analysis of the obtained responses. The development of the scoring system and coding of individual answers should be performed by multiple "experts" who are knowledgeable in the subject field and trained to perform these tasks. High inter-rater agreement in coding responses is necessary to achieve

reliability of scoring. Compared with closed-ended questions, there is a higher risk of participants misunderstanding open-ended questions and providing irrelevant answers or choosing not to answer these questions. For these reasons, most survey questions are closed ended or only partially open ended.

Closed-ended questions list several response options from which the respondent needs to choose. For instance, the question "What is the main stressor in your life?" might be followed by these response categories:

— Finances

— Work

— Relationship with spouse

— Other relationships

— Children

Because closed-ended questions provide a limited number of specific responses, they are much easier to summarize and analyze. These answers are also more relevant and comparable than open-ended responses. However, the participant is forced to choose from alternatives that might not provide the most accurate answer. This problem is addressed in partially open-ended items that provide a set of closed-ended responses (as in closed-ended questions), as well as an open-ended option, such as

— Other (Specify): _____

### *Types of Closed-Ended Questions*

Depending on the format of the response categories, closed-ended questions might be dichotomous, nominal, ordinal, continuous, or rank ordered. Dichotomous questions offer only two possible answers (e.g., yes/no, true/false). Nominal questions include three or more response options that cannot be ordered. For example, the question on the main stressor given in the previous example is a nominal question. In contrast, ordinal items include three or more response categories that can be ordered on a continuum ranging from low to high levels. A set of such ordered responses is typically called a rating scale. Examples include the following:

How stressful is your job?

— Not at all

— Somewhat

— Very much

I often fidget with my hands and feet.

— Strongly disagree

— Disagree

— Agree

— Strongly agree

Rating scales sometimes involve numeric values attached to the descriptors, as in the following example:

How would you rate your health?

1 2 3 4 5

Poor Fair Good Very good Excellent

Continuous questions are a variation of rating scales. Instead of several specific answer options, participants place their answer anywhere on a solid line with descriptors (called anchors) on either side:

Strongly disagree Strongly Agree

Finally, rank-ordered questions ask respondents to rank a set of responses based on their preference or importance. An example would be:

Rank the following leisure activities in order of your preference (1 = most preferred and 5 = least preferred)

— Reading

— Watching TV

— Playing sports, exercising

— Traveling

— Going out

The most important feature of rank-ordered questions is that they force participants to discriminate among the categories. In contrast, when each option is independently rated (e.g., How much do you enjoy reading? How much do you enjoy

watching TV?), respondents might provide the same rating for some or all categories. Although ranking questions are cumbersome to analyze statistically, they might represent a good choice when the order of several alternatives is of main interest.

## Tips on Creating a Survey

The following formatting guidelines will improve the clarity of a survey and result in more complete and accurate answers:

1. Use questions that are short and specific (vs. vague or too general) and include simple rather than complex words. This will increase the likelihood that respondents will correctly understand your questions.
2. Ask only one question in each item. Asking about multiple things will confuse participants and make interpretation of responses difficult. For instance, instead of asking "Do you think parents and teachers should provide civics education for children?," ask two separate questions, one about parents and one about teachers.
3. Visually separate questions from answers. This is typically done by using boldface font for the question and regular font for the answers.
4. List all response options in a single column, with the check boxes, blank lines, or numbers to be checked or circled ordered in a parallel column on the same side of the response options (left or right) throughout the survey. This gives all responses equal "weight" and makes it easier to find and record the best answers.
5. Order questions so that related items are presented together as a group, and that topics are logically ordered throughout the survey. Good organization keeps participants focused on each topic and makes the questionnaire easier and more enjoyable to complete.
6. Begin the questionnaire with easy to understand, noncontroversial questions. Place more personal and sensitive questions (e.g., about sexual behavior) later in the survey. Doing so will reduce the likelihood of respondents abandoning the questionnaire early or not completing the more sensitive questions.
7. Pilot your survey on a small group of individuals to ensure correct understanding and interpretation of the questions.

## Methods of Administration

Surveys can be administered to participants in multiple ways, including through the mail, telephone, Internet, or in person, either individually or in a group setting. The advantages and disadvantages of each method are summarized in the next section.

### *Mail*

After receiving the survey by mail or directly from the researcher, participants complete it at their convenience and return it by mail. This method of administration is convenient and low in cost, and it allows sampling across a wide geographic area. However, respondents cannot ask for clarification and the researcher has no control over who completed the survey and under what conditions. Other major problems of mail surveys include very low response rates and a high risk of nonresponse bias. This bias arises when those who return the survey are significantly different from those who do not respond. As a result, the results from the survey are not representative of the larger population that was surveyed. To increase the return rate of mail surveys, the following two strategies have been shown most effective:

1. Multiple mailings, including a notice preceding the survey, the actual survey, a thank you/reminder letter, and a replacement survey. The mailings should explain the importance of the survey and the respondent's participation and assure confidentiality of answers. Telephone numbers can also be provided that the participants can call if they have questions.
2. Small monetary incentive (e.g., $1) included with the questionnaire. Interestingly, including a small incentive with the survey seems more effective than the promise of a larger incentive contingent on the completion of the survey.

### *Telephone*

Surveys can be administered by telephone, either using live interviewers or interactive voice response (IVR) systems in which participants respond to prerecorded questions using touch-tone phones or by providing spoken answers. Live interviewers are more expensive but can help clarify

questions or response options. However, interviewer bias becomes a potential problem. This bias arises when interviewers' behavior (e.g., how they ask questions or how they react to answers) affects participants' responses. To prevent such bias, interviewers need to be well trained and monitored throughout the data-collection process to pose questions in a neutral manner and in the same way for each participant. Telephone surveys are more cost efficient than in-person interviews, especially when participants are dispersed across a wide geographic area, and they are thus well suited for surveying national or international samples. However, a sampling bias is likely because not all individuals have a phone and many who do screen their calls using caller ID and answering machines. Given the high rates of unsolicited phone calls in recent years, state and federal laws might prohibit such calls and many respondents might be unwilling to answer questions over the phone. However, if reached, individuals might be more likely to respond to a telephone survey compared with a mailed questionnaire.

### *Internet*

Internet surveys can be posted on a website or distributed by e-mail. Compared with e-mail surveys, Web-based questionnaires can be more complex and interactive. Several companies offer survey services on the Internet. Although the initial setup costs can be quite high, the data collection itself is inexpensive. The main advantage of Internet survey is the ability to reach a large number of participants across virtually unlimited geographic area. Additionally, most Web-based surveys automatically create data sets that can be readily analyzed, thus eliminating the tedious and error-prone step of manual data entry. In contrast, samples obtained with Internet surveys tend to be nonrepresentative of the general population, as not all segments of the population have access to the Internet (e.g., the samples might overrepresent younger and more educated individuals). As with mail surveys, the researcher has little control over who completes the survey and under what conditions. Additionally, unless certain protections are taken (e.g., use of passwords), some individuals might complete the survey multiple times, thus biasing the results.

### *Face to Face*

In face-to-face interviews, live interviewers meet with participants individually in a research office, participants' home, or another location. Although this administration method is expensive, it gives researchers more control over the interview process and typically results in higher participation and completion rates, even for long and complex surveys. Participants are able to ask for clarification, thus reducing bias resulting from misunderstanding or misinterpretation of questions. Additionally, this is the only method of administering a survey where interviewers can react to participants' nonverbal communication (e.g., facial expression suggesting confusion) and adjust the interview accordingly (e.g., explain the question or provide assurance). Using face-to-face interviews, structured questionnaires can be easily supplemented by unstructured or semi-structured interviews collecting more in-depth information. However, interviewer bias is a potential problem, so extensive training and monitoring of the interviewers is essential. To obtain valid information, it is important that face-to-face interviews be conducted privately, because the presence of others (e.g., family members, friends) might result in biased answers, especially for sensitive questions. In recent years, many complex surveys have used computer-assisted personal interviewing (CAPI) to make the data-collection process more efficient. Using CAPI reduces the likelihood of interviewer errors, allows for direct programming of complex skip patterns, and automatically records data, thus eliminating the need for manual data entry. Sensitive questions might be asked without the interviewer's involvement using audio computer-assisted self-interview (ACASI) to maximize the validity of the responses.

### *Group Administered*

Typically used in classroom or other group settings, group administration involves many participants completing the questionnaires by themselves at the same time. This method is fast and easy and typically results in high response rates, although a perceived pressure to participate is a concern that should be addressed. The researcher is typically present to answer any questions, and respondents should be encouraged to ask for

clarifications if needed. In a group setting, extra care needs to be taken to ensure confidentiality (e.g., by seating respondents apart from one another or using alternative forms for adjacent individuals). A potential problem is that some participants might not answer as truthfully as when asked in a face-to-face interview.

### Advantages and Disadvantages of Surveys

Surveys represent an easy and efficient way to collect information from a large number of individuals. It is the methodology of choice when the information of interest cannot be obtained from other sources (e.g., questions on personal beliefs) or would be too difficult or costly to ascertain through other means (e.g., from official records). Survey methodology is flexible and can be used to study a broad range of phenomena, including people's behavior, attitudes, beliefs, and emotions. Questions can be tailored to the research topic and can use a variety of formats and response options. The use of standardized, closed-ended questions makes participants' answers easily comparable and quantifiable. Researchers can choose from several administration methods depending on available resources and the type of research and population under study.

In contrast, survey research suffers from many limitations. Respondents might not know the answers to the questions posed or they might interpret the questions and response options differently. For instance, people might be asked to judge certain actions of local politicians of which they have very little knowledge. A given statement (e.g., "The mayor is an effective leader") might have different meaning for different individuals, depending on their beliefs of what characteristics make an effective leader. Individual participants might also have different thresholds for the distinction between "agree" and "strongly agree" or "disagree" and "strongly disagree." Additionally, participants might not be motivated or willing to give true answers, especially to sensitive questions. Even if they are adequately motivated, they might fail to provide valid answers because of inaccurate self-perceptions or biased recall of past events from memory. Some individuals also adopt a *response set*, which is a tendency to provide affirmative or disconfirming answer regardless of the question content (called *yay- saying* and *nay- saying*). The best protection against such response sets is to vary the direction of the response options across questions (e.g., start with "strongly agree" on some and "strongly disagree" on others). The order of questions might also affect participants' responses. Such order effects can be analyzed and adjusted for statistically if the order of questions is varied across individuals. A final problem common in survey research (and pervasive across all human subject research) is that of representativeness. The generalizability of research results is a direct function of how representative the sample is of the sampled population. Even if a solid sampling strategy is used, individuals who consent to participate might differ from those who decline. Consequently, the results of the research might only apply to people similar to those who participated, and not to those similar to the refusers. Many of these limitations can be addressed or minimized by careful survey design, extensive pilot testing, sound sampling strategies, and allocation of resources to maximize response rates and validity of the obtained data.

*Sylvie Mrug*

***See also*** Nonexperimental Design; Quantitative Research; Reliability; Sampling; "Technique for the Measurement of Attitudes, A"; Validity of Measurement; Volunteer Bias

#### Further Readings

Alreck, P. L., & Settle, R. B. (2003). *The survey research handbook* (3rd ed.). New York: McGraw-Hill.

Dillman, D. A. (2000). *Mail and Internet surveys: The tailored design method* (2nd ed.). New York: Wiley.

Fowler, F. J. (1995). *Improving survey questions: Design and evaluation.* Thousand Oaks, CA: Sage.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology.* Hoboken, NJ: Wiley.

## Survival Analysis

Survival analysis, generally speaking, is the modeling of time-to-event data. Traditionally, survival analysis has been used to study the survival of

biological organisms, including human beings and laboratory animals. In this instance, survival analysis seeks to determine the amount of time until the death of the organism. The death is considered the event, and the time until the event is what is modeled. In such instances, survival analysis can determine the proportion of organisms alive at a certain point in time. Furthermore, factors to increase (or shorten) the life span of the organism can be determined. Survival analysis has been applied to many other fields to model data that are not specifically the "life" and "death" of an organism. For instance, survival analysis has been used in engineering and manufacturing to determine the life span of a mechanical process. Death, in this case, is the failure of the mechanism. The social sciences have applied survival analysis to determine the time to events such as marriage, childbirth, and so on. As such, survival analysis is a tool that can be used in a variety of situations, where what is of interest is the time until a particular event can happen.

Regardless of the application, what is central in survival analysis is that the event, be it the death of an organism, the failure of a machine, or the failure of a marriage, is unambiguous. Death is clearly unambiguous, as is marriage. However, the failure of a machine is more ambiguous as failure could be considered partial. Therefore, it is important to define the event in clear terms, and in a way that it is measurable in an absolute way and not open to interpretation. Furthermore, it is assumed that the event can only happen once to a member of the population. Again, in the case of the death of an organism, this is clear. In the case where a machine fails, it can be less clear, depending on the nature of the definition of failure. Many different parts of the machine could fail, and therefore a clear definition of the event should be determined so that the event can only happen once. In the examples of social sciences, this assumption makes it obvious that saying the time to marriage is insufficient; certainly people get married more than once in today's society. However, the event could be defined as first marriage. Doing so also renders the research question more interesting: Given that people marry multiple times, considering the time until marriage in general might be less informative than the time until one gets married for the first time. Precision in the definition of the event can lead to making the second assumption more tenable.

To conduct a survival analysis it is essential to understand the concept of censored observations. Censored observations include observations for which the event has not yet occurred. There are other types of censored observations, but the case where the censorship is caused by not having experienced the event is simplest, and as such will be used for the illustration here. Consider the case where the survival of an organism is of interest. If all the organisms in our sample have died, then looking for a way to predict the survival of an organism is simple: compute the median survival time. This will give us a good indication of the survival time of the organism. (Note that the mean survival time is not likely to be as useful as the median, because survival time data are rarely normally distributed and often are skewed.) In these cases, there is no need for advanced statistical tools. However, if some of the organisms have died while others have not, then a predictive model can be useful. It is important to note, however, that the reason that some organisms have not died (i.e., experienced the event) is because of a limitation of time in the study and that given enough time all organisms would die.

## The Model

The underlying model for survival analysis is the survival function. The survival function, $S(t)$, is given by

$$S(t) = P(T > t), \tag{1}$$

where $P$ denotes the probability, $T$ is the random variable that denotes the survival time, and $t$ is some fixed time. So, conceptually, the survival function is the probability that the survival time is greater than some fixed time. Typically, it is assumed that $S(0) = 1$, which means that the probability that the survival is greater than 0 is 1. There are cases where this is not reasonable, however, and it is not a necessary condition.

The survival function must be nonincreasing, meaning that as $t$ increases, the probability of survival decreases or stays the same; the probability of survival cannot increase over time. Furthermore, it is assumed that as $t$ gets increasingly large,

$S(t)$ approaches zero. This would not be the case where eternal life is possible, but generally speaking, is assumed to be true. In addition to the survival function, survival analysis is also concerned with the hazard function.

The hazard function is the instantaneous rate at which events occur, given no previous events. Conceptually that means

$$h(t) = P(\text{Survive in } [t, t+\Delta t) | \text{Survive past } t)]. \tag{2}$$

Formalizing the expression in Equation 2 yields

$$\begin{aligned} h(t) &= \frac{P(t \le T < t+\Delta t)}{P(T > t)} \\ &= \frac{P(t \le T < t+\Delta t | T > t)}{\Delta t} \\ &= -\frac{1}{S(t)} \frac{S(t+\Delta t) - S(t)}{\Delta t}. \end{aligned} \tag{3}$$

Taking the limit of Equation 3 yields

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{P(t \le T < t+\Delta t | T > t)}{\Delta t} \\ &= -\frac{1}{S(t)} S'(t) = -\frac{d[\ln(S(t))]}{dt}, \end{aligned} \tag{4}$$

which is equivalent to

$$S(t) = e^{-\int_o^t h(u)du}. \tag{5}$$

The cumulative hazard function, which is denoted $H(t)$, provides the accumulated risk up to a specified time and is given by

$$H(t) = \int_0^t h(u)du. \tag{6}$$

It is worth noting the difference between the hazard function and the survival function. The survival function, on the one hand, provides the probability of surviving longer than time $t$, that is, $P(T > t)$ for some value of $t$. The hazard function, on the other hand, provides the probability of dying at time $t$ exactly.

## Estimation

The goal of survival analysis is to estimate $S(t)$, the survival function, and $H(t)$, the hazard function. The estimation of $S(t)$ and $H(t)$ depends on the assumptions that are reasonable to make for the analysis. If it is reasonable to assume that all subjects conform to the same survival function, then estimating $S(t)$ is fairly straightforward, and is the focus of this document. $S(t)$ can be estimated either using a parametric survival function or a nonparametric survival function. The nonparametric approach requires fewer assumptions and is estimated based on the empirical data, whereas the parametric approach naturally relies on choosing a distribution that best describes the data, and as such, the fit of the model chosen to the data becomes an issue. Nonparametric methods are usually preferred for their flexibility in modeling the data. As such, a discussion of the nonparametric approach will be given first, followed by the parametric approach.

### *Nonparametric Approach*

In the nonparametric approach, the empirical cumulative distribution of the survival times is used to estimate the survival function. The most popular approach of nonparametric estimation of the survival function is the Kaplan-Meier approach. To understand the Kaplan-Meier approximation, it is necessary to understand some terms. Consider the case where there are $x$ subjects. Let $t_j$ denote the time of death of subject $j$, $d_j$ denote the number of deaths that have occurred in the time interval $[t_{j-1}, t_j]$, and $n_j$ denote the number of subjects at risk of dying at time $t_j$. To compute the Kaplan-Meier survival function, the following formula is used for each value of $t$ in the interval $t \in [t_j, t_{j+1})$. The function is given by

$$S(t) = \prod_{i=1}^{j} \left[1 - \frac{d_i}{n_i}\right], j = 1, 2, 3, \ldots, \tag{7}$$

where $d_i$ and $n_i$ are defined previously. The resulting function is a step function, with a step at each $t_j$. This is best understood using an example.

**Example**

Consider the hypothetical example of testing a new drug that is designed to treat cancer. Two groups are chosen for the study: a treatment group that received the new drug and a control group that did not. The number of days that each patient survived in the trial was recorded. There were 12 people in the treatment group and 13 people in the control group. The study lasted for 6 months or 180 days. Ten subjects died in each group by the end of the study. The number of days of survival for the treatment group was: 40, 45, 53, 72, 145, 175, 176, 178, 179, 180, and 180. One participant was living at the end of the study and one had dropped out after 40 days. These cases would represent censored data, as their survival times are unknown. The number of days of survival for the control group was: 10, 11, 12, 15, 20, 25, 34, 65, 72, and 169. Of the remaining three participants, two were living at the end of the study and one dropped out after 100 days. These three cases are considered censored data as well.

For the two groups, the Kaplan-Meier statistics can be computed and are provided in Table 1.

The Kaplan-Meier estimate for $S(t)$ can be computed for each of the $t_j$ as follows:

$$\begin{aligned} S(t_1) &= S(40) = \prod_{i=1}^{1}\left[1-\frac{0}{8}\right] = 1.0 \\ S(t_2) &= S(45) = \prod_{i=1}^{2}\left[1-\frac{d_i}{n_i}\right] = \left[1-\frac{0}{8}\right]\left[1-\frac{1}{7}\right] \\ &= .857, \end{aligned} \tag{8}$$

and so forth.

The cumulative hazard function, $H(t)$, is typically estimated using the following approach of Arthur V. Peterson:

**Table 1** Kaplan-Meier Statistics Control and Treatment Groups

| *Time of death ($t_j$)* | *Number at risk ($n_j$)* | *Number died ($d_j$)* | *Censored (0=no, 1=yes)* |
|---|---|---|---|
| *Group: Treatment* | | | |
| 40 | 12 | 0 | 1 |
| 45 | 11 | 1 | 0 |
| 53 | 10 | 1 | 0 |
| 72 | 9 | 1 | 0 |
| 145 | 8 | 1 | 0 |
| 175 | 7 | 1 | 0 |
| 176 | 6 | 1 | 0 |
| 178 | 5 | 1 | 0 |
| 179 | 4 | 1 | 0 |
| 180 | 3 | 2 | 1 |
| *Group: Control* | | | |
| 10 | 13 | 1 | 0 |
| 11 | 12 | 1 | 0 |
| 12 | 11 | 1 | 0 |
| 15 | 10 | 1 | 0 |
| 20 | 9 | 1 | 0 |
| 25 | 8 | 1 | 0 |
| 34 | 7 | 1 | 0 |
| 65 | 6 | 1 | 0 |
| 72 | 5 | 1 | 0 |
| 100 | 4 | 0 | 1 |
| 169 | 3 | 1 | 0 |
| 180 | 2 | 0 | 1 |

**Table 2** Values of $S(t)$ for Subjects in Both Groups

| *Time of death ($t_i$)* | *S(t)* | *H(t) (Petersen)* | *H(t) Nelson-Aalen* |
|---|---|---|---|
| *Group 1* | | | |
| 40 | 1.00 | 0.00 | 0.00 |
| 45 | 0.91 | 0.10 | 0.09 |
| 53 | 0.82 | 0.20 | 0.19 |
| 72 | 0.73 | 0.32 | 0.30 |
| 145 | 0.64 | 0.45 | 0.43 |
| 175 | 0.55 | 0.61 | 0.57 |
| 176 | 0.45 | 0.79 | 0.74 |
| 178 | 0.36 | 1.01 | 0.94 |
| 179 | 0.27 | 1.30 | 1.19 |
| 180 | 0.09 | 2.40 | 1.85 |
| *Group 2* | | | |
| 10 | 1.00 | 0.00 | 0.08 |
| 11 | 0.85 | 0.17 | 0.16 |
| 12 | 0.77 | 0.26 | 0.25 |
| 15 | 0.69 | 0.37 | 0.35 |
| 20 | 0.62 | 0.49 | 0.46 |
| 25 | 0.54 | 0.62 | 0.59 |
| 34 | 0.46 | 0.77 | 0.73 |
| 65 | 0.38 | 0.96 | 0.90 |
| 72 | 0.31 | 1.18 | 1.10 |
| 100 | 0.31 | 1.18 | 1.10 |
| 169 | 0.21 | 1.58 | 1.43 |
| 180 | 0.21 | 1.58 | 1.43 |

$$H(t) = -\ln[S(t)] \tag{9}$$

However, some software packages use the Nelson-Aalen estimate given as follows:

$$H(t) = \sum_{t_i \le t} \frac{d_i}{n_i} \tag{10}$$

The Nelson-Aalen approach will always lead to a slightly lower number; however, there is no theoretical advantage to choose one method over the other.

Both the Peterson and Nelson-Aalen approaches were applied to the data in the previous example, and the values for $S(t)$ and $H(t)$ for the subjects in both groups are provided in Table 2.

Because the goal is to compare the survival rates of the two groups, it is also informative to look at the mean and median survival times for the two groups. The median survival time is usually preferred, as the survival data tend not to be symmetric, but skewed. Table 3 provides the mean and median survival times, along with 95% confidence intervals for each of the values.

**Table 3** Mean and Median Survival Rates With 95% Confidence Intervals

| *Group* | *Estimate* | *Lower Limit* | *Upper Limit* |
|---|---|---|---|
| *MEAN* | | | |
| Treatment | 142.09 | 108.77 | 175.42 |
| Control | 74.56 | 36.29 | 112.83 |
| *MEDIAN* | | | |
| Treatment | 176.00 | 140.40 | 211.61 |
| Control | 34.00 | 0.00 | 86.85 |

As can be observed by looking at Table 3, the mean and median survival times are much higher for the treatment group relative to the control group. Furthermore, the confidence intervals for the median survival time do not overlap, indicating that the differences are statistically significant. Therefore, the treatment seems to lead to greater survival than no treatment.

### *Parametric Approach*

As indicated by the name parametric, the parametric approach requires the specification of a distribution that best describes the empirical data. The idea is to find a continuous function that best approximates the empirical survival function. As noted in the previous section, the empirical survival function is a noncontinuous step function. The parametric function offers the advantages of being a continuous function with known moments of the distribution and known distribution form, which allows for ease in estimating the quantiles of the distribution and estimating the expected survival time. Furthermore, if the model is appropriate for the data, then a more precise estimate of the survival function is provided than that obtained with the Kaplan-Meier approximation. Popular distributions used to approximate the empirical function are the Weibull, exponential, log-normal, and log-logistic models. In using one of these distributions, the parameters of the distribution are estimated from the data. The shape of the survival function is often the by-product of choosing the shape of the hazard function, as the survival function is

a function of the hazard function, as is evident from Equation 1. For example, a constant hazard function would have the form

$$h(t) \equiv a > 0. \tag{11}$$

In this equation, $a$ is a constant and can take the value of any real number. If the hazard function is of this form, then the corresponding function for $S(t)$ would be given by

$$S(t) = e^{-\int_0^t a du} = e^{-at}, \tag{12}$$

which is clearly the exponential function. It is not very realistic for the hazard function to be a constant function, as that would imply that the probability of dying at a specified time $t$ is constant across all values of $t$. Therefore, a more flexible set of distributions is usually specified.

The two parameter exponential distribution is commonly used as a hazard function. The form of this function is given by

$$h(t) = \alpha\beta t^{\beta-1}. \tag{13}$$

This provides a broad class of distributions as it has a scale parameter $\alpha$ and a shape parameter $\beta$. If both $\alpha, \beta > 0$, then the resulting survival function is given by

$$S(t) = e^{-\alpha t^{\beta}}, \tag{14}$$

which is the Weibull distribution and has broad applications in many fields.

Different choices of the hazard function will yield different survival functions. The choice of the distribution will be dependent on the nature of the data.

*Lisa A. Keller*

*See also* Distribution; Mean; Median; Nonparametric Statistics; Observations; Parametric Statistics; Weibull Distribution

**Further Readings**

Aalen, O. O. (1978). Non parametric inference for a family of counting processes. *Annals of Statistics, 6,* 701–726.

Fischer, I. (2006). *Kaplan-Meier formula*. Unpublished manuscript.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association, 53,* 457–481.

Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics, 14,* 945–966.

Peterson, A. V., Jr. (1977). Expressing the Kaplan-Meier estimator as a function of empirical subsurvival functions. *Journal of the American Statistical Association, 72,* 854–858.

# SYSTAT

SYSTAT was first created in the late 1970s by a Chicago psychology professor named Leland Wilkinson. It was the first statistical program that ran on a Macintosh version with DOS and UNIX capabilities and was one of the first programs to run on Windows; it quickly became a prominent statistical analysis and graphics tool in the scientific community. In 1995, SYSTAT was sold to SPSS, an IBM® company, ® formerly called PASW® Statistics, which in 2002 sold it to Cranes Software International, based in Bangalore, India. SYSTAT is currently represented in 38 countries and is managed from Chicago, Illinois. SYSTAT prides itself for being customer oriented by creating affordable academic packages, including a free student version called MYSTAT. Both SYSTAT and MYSTAT are available for download on from the SYSTAT website.

SYSTAT version 12 is available on Windows 98 (or later) and requires a minimum of 128 MB RAM and 220 MB of storage. SYSTAT comes equipped with a vast amount of manuals (in PDF form), including a 2,200-page electronic textbook on Statistics I, II, III, and IV, as well as detailed examples, language references, graphics, and so on. The SYSTAT website also provides a toll-free telephone number and an e-mail link for any technical support issues one might have.

The commands in the SYSTAT task bar are set up in a format similar to SPSS, where most analytical tools are located under the "analyze" option. There, the user will find an option to run basic statistics, which include descriptive statistics (mean,

median, mode, range, *N,* sum, variance, skewness, Kurtosis, etc.); confidence intervals; *hypothesis testing,* which includes comparing the means (one-sample *z* test, two-sample *z* test, *t* test for single sample, paired sample *t* test, two-sample *t* test, and the Poisson test); homogeneity of variance tests (single and multiple); correlation (simple, canonical, and Cronbach's alpha); and *proportion.* Also under basic statistics one can find the options for resampling (bootstrap, without replacement, and jackknife), as well as *N-* and *P-tiles.* Regression analyses include linear (stepwise, least-squares, Bayesian, and ridge), as well as nonlinear, robust, logistic, two-stage least squares, and probit.

Univariate analyses of variance include analysis of variance (ANOVA), analysis of covariance (ANCOVA), and Hotelling's $T^2$. ANOVA tests include one- and two-way ANOVAs and repeated measures ANOVA. Available post hoc tests include Bonferroni, Tukey-Kramer Honestly Significant Difference (HSD), Scheffé, and Fisher's least significant difference. ANOVA designs include nested, unbalanced, randomized block, complete block, mixed model, split plot, Latin square, crossover, and changeover. Multivariate analyses include multivariate analysis of the variance (MANOVA) and multivariate analysis of covariance (MANCOVA) with the options to run discriminant, cluster, or factor analysis.

Nonparametric tests available on SYSTAT 12 include Kruskal-Wallis, Mann–Whitney *U* test, sign test, Wilcoxon, Friedman, Quade, and two-sample Kolmogorov–Smirnov tests. In addition, SYSTAT 12 allows the user to check the normality of assumptions in ANOVA, regression, and general linear model (GLM) using the Kolmogorov–Smirnov (one sample), Shapiro-Wilk (two or more samples), or the Anderson-Darling (repeated measures) tests for normality.

SYSTAT offers a wide range of graphs and plots including histograms, bar, line, dot, profile, pyramid, density plots, pie charts, scatter and probability plots, matrices, maps, and function plots. The graphics PDF found on the start page of SYSTAT 12 screen offers a 528-page electronic textbook about the available graphics, how they are used in statistics, where to find them in SYSTAT, and how to interpret them. In addition, SYSTAT offers three-dimensional charts options for line, pyramid, bar, and histogram charts.

SYSTAT is compatible with many other statistical programs and formats. It can open files saved in Microsoft Excel, IBM® SPSS® (formerly PASW Statistics), SAS, Statistica, Stata, Minitab, S-Plus, JMP, Dbase, ASCII, BMDP, and many others. SYSTAT also has the capacity to save its outputs as any of the previously listed formats.

SYSTAT user interface can be divided into three categories: viewspace, workspace, and commandspace. The viewspace includes the startpage, data editor, variable editor, graph editor, and the output editor. The startpage includes recently opened documents, manuals in PDF form, themes, and a scratchpad, where one can write notes about the current document. The data editor allows one to enter the variables as numeric or string (words). The variable editor page allows one to name and label the variables, assign value, category, choose the number of characters (12 is the maximum) and decimals, choose display type (normal, exponential, or date and time), edit formats, and enter comments. The graph editor allows users to edit graphs, insert text, change font, color, and fill components, edit axes, legends, symbols, rotate the figures, as well as zoom in or out on the objects. The output editor allows one to collapse and expand the given output.

The workspace includes the output organizer, examples, and the dynamic explorer. The output organizer tool allows the user to navigate through the data, as well as copy, cut, paste, and delete desired portions of the output. The examples tool is used for executing commands provided in the user manual with a click of a mouse. The dynamic explorer tool allows the users to rotate and animate three-dimensional figures and graphs.

The commandspace is divided into three categories as well, to include interactive, batch, and log. The interactive tab allows the user to work in the interactive mode, where the user can control each function individually. The batch tab allows for the batch mode, where one can open, submit, edit, or run a large amount of commands at once. The log tab allows the user to examine the commands that were executed during the session; it is particularly useful when using the batch mode.

SYSTAT allows the user to have full control of the user interface. In fact, the user can resize, reposition, and rename the provided tabs. In addition,

the user is given the option of having multiple tabs open at once. Moreover, the users have the option to assign a set of commands to a toolbar button to speed up the operation analysis; resize, reorganize, and hide windows; define settings for data and outputs; specify file locations, and add menu items for most frequently used commands.

SYSTAT is user friendly and compatible with many existing statistical programs. It provides a vast amount of electronic manuals in addition to online support. It provides applications for psychology, sociology, engineering, biology, chemistry, medical research, statistics, genetics, environmental sciences, astronomy, anthropology, and toxicology fields. SYSTAT is easily customizable to fit the needs of its consumer and continues to grow everyday.

*Janina L. Scarlet*

***See also*** R; SAS; SPSS; Statistica

**Further Readings**

SYSTAT 12. (2007). *Getting started.* Chicago: SYSTAT Software.

Yang, M. (2003). *A review on SYSTAT (10) on random effects modeling.* Retrieved January 10, 2009, from http://www.cmm.bristol.ac.uk/learning-training/multilevel-m-software/reviewsystat.pdf

Yurkiewicz, J. (2005). *SYSTAT 11C: comprehensive statistical software and great graphics:* Retrieved January 10, 2009, from http://www.lionhrtpub.com/orms/orms-2-05/swr.html

**Websites**

SYSTAT: http://www.systat.com

# Systematic Error

Systematic error is any error that has a consistent effect. Systematic error results from consistent but inaccurate responses. Using the example of a weighing machine, readings that are consistently off in one direction reflect systematic error, although additional nuances will be subsequently discussed. The causes of systematic error could include leading or biased questions, which are often aspects of the measurement process that typically cause respondents to be unwilling to provide an accurate response. Being unwilling to provide an accurate response, respondents might provide a response that is inaccurate yet consistent. For example, with a leading question, respondents might consistently provide a response that is more acceptable. With a question requiring estimation of the amount of beer consumed, respondents might systematically underestimate the true value if they want to downplay their amount of drinking. This entry discusses several types of systematic error.

## Additive Versus Correlational Systematic Error

One way of visualizing systematic error is in terms of a thermometer or a weighing machine that consistently deviates from the true value in a specific direction by a constant sum, say, as a result of an error in calibrating the zero point on the device. Such an error is a type of systematic error called *additive systematic error.* Additive systematic error inflates or deflates responses by a constant magnitude and might result in reduced correlations with other items. It should be noted that additive systematic error can be constant across responses and therefore has no effect on relationships, or it can be partial in the sense that the additive effect deflates or inflates responses to one end of the scale and restricts variance. Such additive error could be caused by several factors, such as leading questions, interviewer bias, unbalanced response categories, consistently lenient or stringent ratings resulting from wording or other factors, or a tendency to agree or disagree. Factors that cause responses to be consistently off in one direction across respondents lead to additive systematic error.

A more problematic form of systematic error is *correlational systematic error,* which consistently deflates or inflates the relationship between two or more variables. Correlational error occurs when individual responses vary consistently to different degrees over and above true differences in the construct being measured; that is, it is a result of different individuals responding in consistently different ways over and above true differences in the construct. Using the weighing machine example, if readings are off in a certain direction and

also in proportion to somebody's weight, then that is an example of correlational error (e.g., if the weighing machine shows an additional 5 pounds for a 100-pound person and an additional 10 pounds for a 200-pound person). An item might have correlational systematic error resulting from a common method, such as extreme response anchors (e.g., hate-love), leading to the use of middle response categories. Correlational systematic error occurs if different individuals interpret and use response categories in consistent but different ways. Examples include an item with correlational systematic error caused by a common method, such as extreme response anchors, leading to the use of middle response categories, or an item with moderate response anchors (e.g., like-dislike), leading to the use of extreme response categories. Correlational systematic errors can be caused by the use of response scales of a similar format across items, including what are referred to in the literature as method factors. For example, a certain set of response categories is employed (say, *very good* to *very bad*) and respondents interpret the categories in certain ways (*very good* means more or less positive for different respondents). In this scenario, the covariance across items will be, at least partially, a result of the method factor, or the use of identical response formats. Similarly, correlational systematic error might arise as a result of other aspects of the research method, such as variation in the interpretation of the instructions and the questions.

Additive systematic error in an administration can have either no effect or a negative effect on stability and test–retest correlations. Additive systematic error is likely to have no effect or to attenuate relationships because of decreased variance. Correlational systematic error can strengthen or weaken observed relationships. Consistent differences across individuals over and above the construct being measured might be positively correlated, negatively correlated, or not correlated with the construct.

## Within-Measure Correlational Systematic Error and Across-Measure Systematic Error

Correlational systematic error can be distinguished within measures versus across measures. The latter, *across-measure systematic error,* refers to correlational systematic error that occurs between measures of different constructs. *Within-measure correlational systematic error* occurs between different items of a measure. It can be the result of responses to items being influenced by the use of a common method or by different trait or specific method factors that might or might not be related to the trait being measured. For instance, all the positively worded items in a scale might be influenced by individual differences in tendencies to agree. Within-measure correlational systematic error can have a positive influence on stability, such as the consistent measurement of a different but stable method factor or construct. An example of within-measure correlational systematic error is halo error in the completion of items within a measure—a tendency to provide similar responses across items that are thought to be related.

Whereas within-measure correlational systematic error occurs between items of a measure, across-measure correlational systematic error occurs across measures of different constructs. Essentially, across-measure correlational systematic error occurs if measures of different constructs are influenced by a trait or a method factor, or by different but related traits or method factors, thus inflating or deflating true correlations between them. A common method factor that affects both measures is a source of across-measure correlational systematic error. For example, the use of the same paper-and-pencil method (that taps, say, a response style of using certain parts of the scale) is likely to inflate correlations. Likewise, if subsequent measures are influenced by earlier ones, then the scales completed first introduce measure. For example, hypothesis guessing might result from responses to the first measure influencing responses to the second measure.

*Madhu Viswanathan*

---

***Note:*** Material in this entry adapted from Viswanathan, M. (2005). *Measurement error and research design.* Thousand Oaks, CA: Sage.

***See also*** Error; Error Rates; Multitrait–Multimethod Matrix; Psychometrics; Random Error; Reliability; Response Bias; Validity of Measurement

**Further Readings**

Viswanathan, M. (2005). *Measurement error and research design*. Thousand Oaks, CA: Sage.

# Systematic Sampling

The techniques of inferential statistics are used to make educated guesses about the unknown characteristics of a population based on the known properties of a sample. There are many different kinds of samples, such as simple random samples, stratified random samples, multistage cluster samples, quota samples, and convenience samples. Another kind of sample is a systematic sample. Such samples differ from each other in terms of quality, availability, and popularity.

Systematic sampling refers to the process used to extract a sample from the population. From an ordered list of the population's $N$ members (people, animals, or things), every $k$th member is selected to be included in the sample, where $k$ is the interval between selected members of the list. The value of $k$ is selected by the researcher to create a sample of a desired size, $n$. To determine $k$, $N$ is divided by $n$.

A simple example might help to clarify how one creates a systematic sample. First, suppose a researcher has a population of 200 members and wants to have a sample of size 25. In this case, $N = 200$, $n = 25$, and $k = \frac{200}{5} = 8$. In this situation, every 8th member on the list of the 200 elements of the population would be included in the sample. The 25 members of the sample would hold positions $8, 16, 24, \ldots,$ and 200 in the list (presuming that the first sampled position on the list is position 8).

This entry describes the process of selecting a systematic sample and discusses the advantages and disadvantages of such a sample.

## Selection Process

### *Need for a Sampling Frame*

To create a systematic sample, there first must be a list of every member of the population. This list is called a *sampling frame*. Examples of sampling frames include the list of all admitted students who enroll in a given university as freshman during the fall of a particular year, the list of all gorillas housed in metropolitan zoos on a particular date in a given country, and the list of all new homes in a particular county that are sold during a given month. In each of these examples, the list of the population's members is the sampling frame.

In some studies, a sample is created by identifying every $k$th person who walks by a researcher who is stationed at a particular spot in a city. After these individuals are identified (and questioned on some topic of interest), with the value of $n$ perhaps predetermined, the responses from the $n$ people are treated as sample data. This kind of sample is not a systematic sample for one simple reason: No sampling frame was created or used. Systematic samples, by definition, require sampling frames.

### *The Starting Point*

Because most lists that define sampling frames are created in an alphabetical, geographical, or time-based manner, it is important to have a randomly determined starting point among the initial $k$ positions on the list. If this is not done, no member of the population has a chance to be included if its position in the sampling frame is smaller than $k$. Clearly, a systematic sample cannot be considered to be random if a portion of the population is prevented from entering the sample.

In the example presented previously (in which $N = 200$ and $n = 25$), $k$ was determined to be 8. Suppose the starting point is randomly determined to be 3. This would mean that members of the population in positions $3, 11, 19, \ldots,$ and 195 in the sampling frame would become members of the sample. This sample would be a true random sample; each member of the population has an equal chance of ending up in the sample, regardless of its location in the sampling frame.

### *Noninteger Values of* k

If the ratio of $N$ to $n$ yields a noninteger value of $k$, then it is not proper to round $k$ to the nearest integer. Doing this will cause one of two problems to occur. If $k$ is rounded down, then one or more members of the population listed at the end of the sampling frame will have no chance of being

included in the sample. If $k$ is rounded up, then the selected sample might end up being smaller than desired.

To illustrate the proper procedure, suppose $N$ is 206 and the desired $n$ is 25. Here, $k$ is equal to 8.24. As before, a random starting point would be identified among the initial 8 members of the sampling frame. (The number 8 came from rounding the computed value of $k$ to the nearest integer.) Next, multiples of the noninteger value of $k$ are added to that starting number and rounded off to determine the 2nd, 3rd, . . . , and last members of the sample. For example, if the starting number for our illustration is 4, then entries into the sample would come from positions:

Startin position is 4. . .

$4 + 8.24 = 12.24$ rounds to 12

$12.24 + 8.24 = 20.48$ rounds to 20

$20.48 + 8.24 = 28.72$ rounds to 29

So, the entries would come from positions 4, 12, 20, 29, . . . , 202 in the sampling frame.

### *Circular Systematic Samples*

A slightly different method of extracting a systematic sample involves thinking of the sampling frame as a circle rather than as a vertical or horizontal line segment. Using this method, the value of $k$ is first determined (and rounded, if necessary, to the nearest integer). Then, beginning with a random starting point between 1 and $N$, every $k$th member of the sampling frame goes into the sample, with this process continuing seamlessly from the "end" to the "beginning" of the sampling frame in those situations where the starting point is not one of the initial $k$ positions in the sampling frame.

Suppose a researcher wants to extract a sample of $n = 25$ from a population in which $N = 418$. Here, $k = 16.72$ and the rounded $k = 17$. If the randomly determined starting point happens to be 391, the circular systematic sample would contain members from the sampling frame located in positions 391, 408, 7, 24, . . . , and 381. Here, position 7 comes from adding 17 to 408 under the stipulation that the string of integers reverts to 1 after reaching 418. This sample (like all circular systematic samples) is truly random because every member of the population has an equal chance of being included in the sample. These samples also have the characteristic of being the desired size, $n$, even in those cases where $N \neq nk$.

## Advantages and Disadvantages

There are two main advantages of systematic samples. First, this kind of sample is easy to create in those situations where (a) $N/n$ is an integer or (b) a circular approach is taken if $N/n$ is not an integer. In these situations, only one random number is needed in the process of selecting the sample. Second, a systematic sample has the appearance of being both good and fair because the sample is created by selecting members from various "sections" of the defined population. This kind of face validity might cause summaries of the sample data to be accepted more readily by lay people who realize that simple random sampling potentially can identify $n$ members of a population who hold consecutive positions in the sampling frame.

The main disadvantage of systematic sampling is the possibility of some form of cycle (i.e., periodicity) in the sampling frame. For example, if every 10th home in a neighborhood is selected to form a systematic sample, and if each of these homes (but not others) has a fire hydrant located directly in front of it, then this sample might be badly biased if a researcher asks the sampled homeowners to indicate how worried they are about having a fire destroy their homes. Or, suppose a systematic sample of $n = 10$ work days is taken from a 50-week period of time, with workers measured on those 10 days in terms of productivity, illness, or some other variable of interest. With $k = 5$, the sampled days could all be Mondays, and the sample data could easily misrepresent the workers' status throughout the 50-week duration of the study.

The problems caused by cycles existing in the sampling frame can be circumvented in one of two ways. One solution is to rearrange randomly the members of the population prior to the time the systematic sample is drawn. The other solution is useful in those situations where it is inconvenient or impossible to rearrange the sampling frame. In

such situations, potential problems caused by known or hidden cycles can be eliminated by (a) randomly selecting a starting point in the existing sampling frame and (b) randomly selecting a member from each of the *k*-sized intervals that are positioned after the starting point.

*Schuyler W. Huck, Shelley Esquivel, and Amy S. Beavers*

***See also*** Cluster Sampling; Convenience Sampling; Inference: Deductive and Inductive; Population; Quota Sampling; Random Assignment; Random Sampling; Random Selection; Sampling; Stratified Sampling

**Further Readings**

Daroga, S., & Chaudhary, F. S. (1986). Systematic random sampling. In *Theory and analysis of sample survey designs* (pp. 81–109). New York: Wiley.

Kish, L. (1967). *Survey sampling*. New York: Wiley.

Murthy, M. K. (1967). *Sampling: Theory and methods*. Calcutta, India: Statistical Publishing Society.

Sampath, S. (2005). *Sampling theory and* methods. Harrow, UK: Alpha Science International Ltd.

Snedecor, G. W., & Cochran, W. G. (1988). Systematic sampling. In *Statistical methods* (pp. 446–447). Malden, MA: Blackwell.

Som, R. K. (1995). *Practical sampling techniques*. Boca Raton, FL: CRC Press.

# T

## "Technique for the Measurement of Attitudes, A"

Often, researchers in the social sciences would like to quantitatively describe or explain people's attitudes or beliefs about an issue. The problem is that attitudes or beliefs—such as "hope," "customer satisfaction," or "conservatism"—are qualitative and cannot be directly measured the way weight, height, or speed can be measured. This is the problem that Rensis Likert and others were up against in the early 20th century, and which Likert addressed in his article "A Technique for the Measurement of Attitudes."

Until 1932, the best procedures available for measuring attitudes were those developed by Louis Thurstone. In Thurstone's approach, as many as 200 experts, called judges, each rated a large number of statements for their favorableness toward a certain position or object. Through a process of elimination, 20 to 30 statements would be retained and ordered along a scale representing the attitude toward the position or object, from negative through neutral to positive. Based on its position on the continuum, each statement was then assigned a scale score. Thereafter, to measure a person's attitude toward that object or issue, he or she would be asked whether they agree with each statement. The person's attitude score was obtained by taking the average scale scores of all the statements that person agreed with.

Likert identified two shortcomings in Thurstone's procedure. First, it required many judges and a long, laborious process. Second, the use of the attitude scales made several statistical assumptions that had not been verified. In his 1932 article, "A Technique for the Measurement of Attitudes," Likert introduced a streamlined approach to improve on Thurstone's methods. His main goal was to simplify the process by creating the scale from participants' responses and thus eliminating the influence of the judging group. Whereas previous methods engaged separate processes for assigning scale scores to statements and attitude scores to individuals, in Likert's study, each statement became a scale in itself, and a person's responses to each statement were assigned a score. These response scores were then combined by using a median or mean to obtain an attitude score.

Although Likert's 1932 article is most noted for the scales that carry his name, the issue he tackled was not purely a matter of simplifying measurement but of determining whether in fact social attitudes could be measured reliably and validly. If so, it should be possible not only to measure a general attitude but also to justify its difference from other attitudes. Likert's 1932 study showed, to his own surprise, that general social attitudes could be measured validly and reliably.

This remainder of this entry describes the development of the Likert scales and examines their shortcomings.

## Development of the Measure

In this article, Likert presents the results from a study conducted under Gardner Murphy to examine five major "attitude areas": international relations, race relations, economic conflict, political conflict, and religion. The participants in the study included about 650 students at universities in the United States. Every participant took a survey twice, about a month apart. The second survey included some items from the first survey in addition to many new items.

To construct the survey, Likert and his colleagues consulted existing surveys, as well as newspaper and magazine editorials and articles, for statements that would effectively distinguish people as sitting at different points along the continua under investigation. Thus, for example, international relations items ranged from anti-internationalism to pro-internationalism. Three kinds of items were presented to participants. Some items, such as "Should there be a national referendum on every war?" were posed as "Yes/?/No" questions. For the second kind of statement, five responses were designed to span the range of attitudes. An example of this kind of question is as follows:

How much military training should we have?

- We need universal compulsory military training.
- We need citizen military training camps and reserve officers training camps but not universal military training.
- We need some facilities for training reserve officers but not as much as at present.
- We need only such military training as is required to maintain our regular army.
- All military training should be abolished.

The third kind of statement included those whose response options ranged from "strongly disapprove" through "disapprove," "neutral," "approve" to "strongly approve." An example of this kind of statement is as follows: "The United States should have the largest military and naval air fleets in the world."

Likert expected that attitudes would be specific, although not completely independent of each other. For instance, for the race-relations attitude, which given the historical period was defined as "Negro-white relations," Likert expected that the white male participants' attitudes toward segregation, toward eating with the Negro, and toward lynching would be independent. The assumption was that attitudes were so specific that it would take a large number of items to capture anything approximating a general attitude. To his surprise, however, clear group factors appeared in the data: "The clear-cut generality of certain attitudes, such as pro-Negro, internationalism, etc., shows that it is precisely in the field of *affiliation with or against* certain social groups that the most definite results are obtained" (pp. 13–14). He based this finding on the high split-half reliabilities (.86–.88) of statements related to each attitude but addressing specific aspects of it, and on the finding that correlations between different attitude scales were much lower (.34–.63).

Another major contribution of the article was the finding that a simple scoring method worked just as well as the more labor-intensive sigma method devised by Edward L. Thorndike. First, Likert applied the sigma method to transform responses to each item into sigma values. Sigma values are similar to standardized ($z$) scores, except that the sigma method assumes that 100% of responses fall between –3 and +3 sigma. An individual's attitude score was the average of the sigma values of his or her responses. He then applied an alternative approach, which he dubbed the "simpler method," in which scores from 1 to 5 were assigned to the options from strongly disapprove to strongly approve, and scores of 2, 3, and 4 were assigned to the "Yes," "?," and "No" responses, respectively. The score for each individual was then calculated as the average of the numerical values of the options he or she had selected. When he examined the individual attitude scores generated by the sigma and simpler methods, he found that they were nearly perfectly correlated ($r_{xy} = .99$).

Likert then compared the results from his study with those obtained using the Thurstone method, and he found that scores derived from the simple method were highly correlated with scores from the Thurstone method, even though different items were used in the two surveys.

This article also introduced substantive psychological findings. In Likert's words, "at least three general or group factors in social attitudes have been discovered within the student populations which participated in our study" (p. 36). "If the

present method has any value at all, it has value as a device for laying bare the general favorableness or unfavorableness which certain individuals and groups display toward other individuals and groups" (p. 39).

Today, researchers commonly use Likert scales in the development of new instruments. For example, if the options on a rating scale are successively ordered, as in Likert scale, then a graded response item or a Rasch family of polytomous item response model can be applied. An item response model expresses a probabilistic relationship between examinee's performance on a test item and the examinee's latent trait.

## Shortcomings

Despite its advantages, Likert scaling does possess several shortcomings. Likert's method for measuring attitudes assumes that participants' responses to items refer wholly to the properties of the attitude (or latent trait) being measured. Other researchers, such as Louis Guttman, have asserted that the rating a person assigns to an item reflects a property of the item as well as a property of the person doing the rating. For instance, if several people were asked to rate the weight of several objects, each rating would be determined by a combination of the object's weight and the particular person's strength. Guttman developed methods for evaluating whether the rating can be used to estimate the magnitude of the person trait.

Another major issue with Likert scales is that they tend to confound two dimensions of an attitude: direction (i.e., positive, neutral, or negative) and intensity (e.g., strongly or mildly), and thus, these scales do not yield unidimensional ordinal scores. In other words, Likert scales directly measure the interaction of direction and intensity and indirectly measure the main effects of direction and intensity. As such, the main effects are inferred from the interaction. This issue is particularly problematic when more than one dimension is present in the response key because there is no way to individuate the different dimensions. Researchers have attempted to address the problem of multiple dimensions by separating the cognitive (content) and affective (intensity) dimensions into two discrete response keys, one that asks participants whether they agree with the statements and the other that asks participants how strongly they feel about the statements.

Likert scales commonly incorporate negatively worded items to circumvent the problem of response-set bias—the tendency of respondents to agree with a series of positively worded items. However, the added cognitive complexity associated with negatively stated items results in lower levels of validity and reliability.

A similar set of issues also might come into play when respondents disagree with positively (agree) worded statements. Although it is probably safe to assume that the level of agreement indicates the degree of the underlying attribute, disagreement is problematic, because respondents might disagree with a statement for any number of reasons aside from the attitude of interest. This issue is thus of particular concern when a respondent's disagreement is taken to indicate the presence of a particular trait.

Additional problems occur with the use of an odd number of response categories because individuals might equate the midpoint option with a "not applicable" response, whereas the researcher interprets the midpoint option as a mid-level-intensity response.

*Michelle M. Riconscente and Isabella Romeo*

***See also*** Guttman Scaling; Interval Scale; Item Response Theory; Likert Scaling; Nominal Scale; Ordinal Scale; Ratio Scale; Thurstone Scaling

### Further Readings

Albaum, G. (1997). The Likert scale revisited: An alternate version. *Journal of the Market Research Society, 39,* 331.

Hodge, D. R., & Gillespie, D. (2003). Phrase completions: An alternative to Likert scales. *Social Work Research, 27,* 45.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychologist, 140,* 5–53.

Massof, R. W. (2004). Likert and Guttman scaling of visual function rating scale questionnaires. *Ophthalmic Epidemiology, 11,* 381–399.

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement, 14,* 59–71.

Roberts, J. S., Laughlin, J. E., & Wedell, D. H. (1999). Validity issues in the Likert and Thurstone approaches

to attitude measurement. *Educational and Psychological Measurement, 59,* 211–233.
Shrigley, R. L., & Koballa, T. R. (1984). Attitude measurement—judging the emotional intensity of Likert-type science attitude statements. *Journal of Research in Science Teaching, 21,* 111–118.
Thurstone, L. L. (1928). Attitude can be measured. *The American Journal of Sociology, 33,* 529–554.

# *Teoria Statistica Delle Classi e Calcolo Delle Probabilità*

Carlo Emilio Bonferroni was an Italian mathematician, who is known in the field of statistical simultaneous inference for the so-called Bonferroni inequalities that Bonferroni described in his book *Teoria Statistica Delle Classi e Calcolo Delle Probabilità* [*Statistical Class Theory and Calculation of Probability*]. This work has made significant contributions to probability theory and statistical inference.

The book, divided into three sections, is a systematic study of probability theory and its applications on set (class) theory. In the introduction, Bonferroni states that the probability of an event is a primitive physic magnitude that does not have a definition; however, it is possible to explain it in some postulate or axiom. The classical postulate assumes that if an event is separated by other incompatible, complementary, and equally possible events, then the probability of that event is the ratio between the number of cases favorable to it and the number of all events possible.

## Relationship Among Probabilities

In the first section, Bonferroni, before enouncing his inequalities, describes the following probabilities and their relationship:

*Simultaneous probabilities:* In a set composed of $m$ objects, if $n$ characteristics $C_1$, $C_2 \ldots \ldots C_n$ are considered and indicated by $m_i$, the number of objects having the characteristic $C_i$, by $m_{ij}$ the number of objects having simultaneously $C_iC_j$ and by $m_{ijh}$ those having $C_iC_jC_h$ etc... then the probabilities are

$$p_i = \frac{m_i}{m}, p_{ij} = \frac{m_{ij}}{m}, p_{ijh} = \frac{m_{ijh}}{m}, \ldots$$

*Contrary probabilities:* $p_{\overline{ij\ldots k}}$ is the contrary probability to $p_{ij\ \ldots k}$, that is the probability of an object without having the characteristics $Ci$, $Cj, \ldots, C_k$. The probability is

$$p_{\overline{ij\ldots k}} = 1 - p_{ij} \ldots k$$

and in more general

$$p_{\overline{12\ldots n}} = 1 - \sum_{i}^{l,n} p_i + \sum_{i,j}^{l,n} p_{ij} - \sum_{i,j,h.}^{l,n} p_{ijh} + \ldots p_{12\ldots n}.$$

If we define

$$S = 1, S_1 = \sum\nolimits_{p_i}, S_2 = \sum\nolimits_{p_{ij}}, S_3 = \sum\nolimits_{p_{ijh}}, \ldots,$$

then it is possible to write

$$p_{\overline{12\ldots n}} = S_0 - S_1 + S_2 - S_3 + \cdots \pm S_n = S_0(1+S)^{-1}.$$

*Exclusive probability:* A probability that an object has some characteristics excluding the others. It is defined by the product between the probabilities of those characteristics taken into consideration and the contrary probabilities.

*r–exact multiplicity:* The multiplicity of an object is defined as the number of characteristics that it possesses, from 0 to $n$.

*r–minimum multiplicity:* An object that possesses at least $r$ characteristics.

*r–maximum multiplicity:* An object that possesses at most $r$ characteristics, that is 0, 1, 2,... or $r$.

*Incompatible characteristics, complementary characteristics:* The characteristics $C_i$ are incompatible when an object has at most one; for $C_iC_j$, $p_{ij} = 0$. They are complementary if an object possesses at least one. The formula is

$$1 - S_1 + S_2 - S_3 \ldots \pm S_n = 0.$$

The characteristics are incompatible and complementary when $1 - S_1 = 0$; that is, $\Sigma p = 1$.

*Identity:* $(1 + S)(1 - P) = 1$.

*Contrary characteristics and duality law:* If $m$ objects are described by their contrary "not $C_i$" characteristics rather than by the characteristics $C_i$, the simultaneous probabilities are $q_i$, $q_{ij}$, $q_{ijh}, \ldots$ and are opposites of the affirmative probabilities $p_i, p_{ij}, p_{ijh}, \ldots$. For the *duality law*, each calculation developed on the affirmative probabilities is reproducible on the contrary probabilities.

$$T_0 = 1, T_1 = \Sigma q_i, T_2 = \Sigma q_{ij}, \ldots.$$
$$q_{\overline{1\ldots n}} = T_0 - T_1 + T_2 - \cdots \pm T_n = T_0(1+T)^{-1}.$$

The duality law allows expression of both affirmative and contrary probability in a double way, through S and T:

$$T_n = 1 - S_1 + S_2 - S_3 \ldots, \pm S_n$$
$$T_{n-1} = n - (n-1)S_1 + (n-2)S_2 - \cdots \pm S_{n-1}$$
$$T_{n-2} = \binom{n}{2} - \binom{n-1}{2}S_1 + \binom{n-2}{2}S_2 - \cdots \pm S_{n-2}$$

*Limits on simultaneous probabilities:* From $q_{12\ldots n} = P_0 = 1 - S_1 + S_2 + S_3 + \cdots \pm S_n$, Bonferroni developed his inequalities:

$$P_0 \le 1, P_0 \ge 1 - S_1, P_0 \le 1 - S_1 + S_2,$$
$$P_0 \ge 1 - S_1 + S_2 - S_3, \text{ etc.}$$

The first inequality is evident, whereas the second inequality as well as the third one can be obtained with the following procedure:

For the second inequality: $q_{12} = (1 - p_1)(1 - p_2) = 1 - S_1 + S_2 \ge 1 - S_1$.

For the third inequality: $q_{123} = (1 - p_1)(1 - p_2)(1 - p_3) = 1 - S_1 + S_2 - S_3 \le 1 - S_1 + S_2$.

Applying the duality law and considering the relationship between T and S, the inequalities are

$$p_{12\ldots n} \le S_r - \binom{n-r}{1}S_{r-1} + \cdots + \binom{n-1}{r}, [r \text{ even}]$$

and

$$p_{12\ldots n} \ge S_r - \binom{n-r}{1}S_{r-1} + \cdots - \binom{n-1}{r}, [r \text{ odd}].$$

These formulas are the basis of Bonferroni correction or adjustment used in the simultaneous statistical inference for multiple comparison tests.

## Dependency and Independency

The second section studies the dependency or independency (total or partial) of characteristics in the classes, proposing an index of dependency:

$$I = \frac{p_{12} - p_1 p_2}{\sqrt{p_1 p_2 q_1 q_2}}.$$

Bonferroni defines a generated class of a probability distribution as that which possesses objects with $n$ characteristics and $r$ multiplicity and its probability is $P_r = p_{12\ldots n(r)}$ (with $r$ multiplicity and $n$ characteristics). There are different generated classes; however, he establishes that these classes do not exist when the characters are totally independent.

Then, the author defines the "capacity" of an object of the class as the ratio between the multiplicity of an object $X_i$ and all the characteristics of that object taken in consideration $x_i = \frac{X_i}{n}$. The author affirms that its variability $\delta$, expressed as $\delta^2 = \frac{1}{n^2}[(\Sigma p_i - \Sigma p_i) + \Sigma p_{ij}]$, is always between 0 and $\frac{1}{2}$ and might reach extreme values; in the case of total binary independence, it will not exceed $\frac{1}{2}\sqrt{n}$.

## General Laws of Probability

The last section treats the laws of probability. In a scheme of probabilities, the author studies the prevalent points (central point as well as maximum and minimum prevalent points) and its properties, analyzing the relationship between these points and the capacity variability. He establishes that when this variability tends to zero, the capacity tends toward the mean capacity.

Finally, he shows that through a set of simple classes, it is possible to obtain an independent total scheme of probability and apply it to the Poisson and Bernoulli scheme.

*Antonio Miceli*

***See also*** Bonferroni Procedure; Multiple Comparison Tests

**Further Readings**

Bonferroni, C. E. (1935). Il calcolo delle assicurazioni su gruppi di teste. In *Scritti in onore di S. Ortu-Carboni.* Genoa, Italy: R. Istituto Superiore di Scienze economiche e commerciali. [Calculation of insurances among groups of assureds. In *Written in honor of S. Ortu-Carboni.* Genoa, Italy: High R Institute of Economic and Commercial Sciences.]

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze, N.8.* [Statistical class theory and calculation of probability. *Publications of High R Institute of Economic and Commercial Sciences of Florence, N.8.*]

# Test

Hypothesis testing is a formal procedure often employed in scientific research to test theories or models. To formulate such a test, usually some theory has been put forward, either because it is believed to be true or because it is to be used as a basis for argument, but it has not been proved yet. In effect, the researcher bets in advance of the experiment that the results will agree with his or her theory and cannot be accounted for by the chance variations that are involved in the sampling. Nowadays, there are at least four different schools of thought on the inferential significance testing: Fisherian, Neyman–Pearson, Bayesian, and likelihood inference. It would be fair to argue that none of these inferential statistical methods is without controversy.

Before discussing these controversies, this entry provides a brief history and describes classical hypothesis testing. Last, this entry discusses future research, including the use of nonparametric tests.

## Brief History

Statistical significance testing has been dated to 1900 with Karl Pearson's publication of his chi-square "goodness-of-fit" test comparing data with a theoretically expected curve. In 1908, William Sealy Gossett (under the pen name of Student) set the stage for the "classical" hypothesis testing. He introduced the notions of the *test statistic* and its *p value* (the *probability* of obtaining a *test statistic* at least as extreme as the one that was actually observed, given that the *null hypothesis* is true). Ronald Aylmer Fisher is credited for overemphasizing the role of tests of significance using 5% and 1% quintiles of test statistics in his 1925 articles. In a series of articles written between 1924 and 1934, Jerzy Neyman and Egon Pearson developed the statistical hypothesis test procedures used in every statistics textbook today.

## Classical Hypothesis Testing

There are two types of statistical hypotheses, the null and alternative, which deal with the numerical value of a specific population parameter. The null hypothesis is generally the opposite of the research hypothesis, which is what the investigator truly believes in and wants to demonstrate. The test is designed to determine whether an alternative hypothesis achieves the required level of statistical significance, which would justify the acceptance of the alternative hypothesis in preference to the null hypothesis. The decision rule is based on the relevant sampling distribution of the test statistic for the population parameter under the null hypothesis and the selected value of the significance level. The actual decision is based on the selected decision rule and the result of a random sample taken from the relevant population. The decision is to reject or fail to reject the null hypothesis. If the result of the test does not correspond with the actual state of nature, then an error has occurred. There are two kinds of error, which are classified as Type I error (when the true hypothesis is wrongly rejected) and Type II error (when a false hypothesis is not rejected), depending on which hypothesis has been incorrectly identified as the true state of nature. There are several statistical tests to choose from, and choosing the right one for a particular set of data can be an overwhelming task. It is well known that both errors' probabilities cannot be minimized simultaneously. The Neyman–Pearson decision-theoretical approach to finding the best test is to assign a small bound on Type I error probability and minimize Type II error probability.

## Hypothesis Testing Controversies

The controversy between Fisher and the Neyman–Pearson school was never effectively resolved, even though the contribution of Neyman, Karl Pearson,

and Egon Pearson, whose views Fisher so vehemently opposed, has become a standard practice. Oddly enough, these two schools of thought came to be perceived as a unified theory in textbooks. Fisher promoted the *p* value for testing in a wide variety of problems, rejecting the null hypothesis when the *p* value is too small. Pearson emphasized that the *p* value is used to judge whether a given hypothesis provides a good agreement with data and not for accepting as true or rejecting as false a hypothesis. Fisher always rejected the core concepts of the Neyman–Pearson approach: the notion of Type II error, the idea of repeated samples, and the acceptance of the alternative hypothesis obtained through the rejection of the null hypothesis. Alternatively, Neyman argued that Fisher had no logical basis for the choice of his test statistics and that it did not obey the frequentist criteria.

Bayesians argue that the testing of a hypothesis has no logical basis and that one should start with a possible alternative hypothesis with known prior probabilities of being true and derive posterior probabilities in the light of observed data. The major criticism of the Bayesian approach is that it is difficult and/or inappropriate to choose a prior distribution for Bayesian testing.

Within the likelihood paradigm, if an event is more probable under hypothesis A than B, then the occurrence of that event is evidence supporting hypothesis A over B, and the likelihood ratio is used to measure the strength of statistical evidence. The likelihood approach is inherently comparative, requiring several hypotheses pitted against each other. Hence, testing becomes a decision procedure for choosing one hypothesis over another.

### Future Research

*Nonparametric tests* are often used in place of their parametric counterparts when certain assumptions about the underlying population are not satisfied. *Permutation tests* are tests in which a reference distribution is obtained by calculating all the possible values of the test statistic under rearrangements of the labels on the observed data points. An important assumption behind the permutation test is that the observations are exchangeable under the null hypothesis. The idea behind *bootstrap tests* is to use bootstrapping to approximate the distribution of a suitable test statistic when the null hypothesis is true.

The major drawback of nonparametric tests is their computational expense. More research in the development of a new, sophisticated algorithm would make their applicability to a wide range of problems more practical as well as facilitate the integration of nonparametric options into the main statistical software packages.

*Tatiyana V. Apanasovich*

***See also*** Alternative Hypotheses; Bootstrapping; Nonparametric Statistics; Null Hypothesis; Parametric Statistics; Type I Error; Type II Error

#### Further Readings

Efron B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Lehmann, E. L. (1986). *Testing statistical hypothesis* (2nd ed.). New York: Springer-Verlag.

## Test–Retest Reliability

Test–retest reliability is one way to assess the consistency of a measure. The reliability of a set of scores is the degree to which the scores result from systemic rather than chance or random factors. Reliability measures the proportion of the variance among scores that are a result of true differences. True differences refer to actual differences, not measured differences. That is, if you are measuring a construct such as depression, some differences in scores will be caused by true differences and some will be caused by error. For example, if 90% of the differences are a result of systematic factors, then the reliability is .90, which indicates that 10% of the variance is based on chance or random factors. Some examples of chance or random errors include scoring errors, carelessness on the part of the respondent (e.g., not clearly marking an answer), and outside distractions on the day the test is administered (e.g., someone talking loudly near the testing room). Determining the exact true score for each subject is not possible; however, reliability can be estimated in several ways. Each method of determining reliability has advantages and disadvantages. This entry describes the ways to measure reliability

and discusses the assumptions and considerations relevant to the application of test–retest reliability.

## Methods

One way to measure reliability is to determine the internal consistency of a measure. If the various components or items of an instrument are measuring the same construct, then the scores on the components or items will tend to covary. That is, if the instrument is all keyed in the same direction, then people that are high on the construct (e.g., extroversion) will tend to answer all the items in one direction, and people low on the construct (e.g., those who are not extroverted) will tend to answer all the items in the opposite direction. Cronbach's alpha coefficient and split-half reliability are estimates of internal consistency reliability.

A second common estimate of reliability examines the stability of the scores. One way to measure the stability of a measure is to compare alternate or parallel forms of the instrument. High-alternate-form coefficients indicate the forms are comparable and reliable. Test–retest reliability is the most common way to measure the stability of a measure over time. It is conceptually and intuitively the simplest approach and one that most closely corresponds to the view of reliability as the consistency or repeatability of a measure. That is, if a researcher has the same people take the same test on more than one occasion (i.e., a first test occasion and a retest occasion), the correlation between each test administration will be the test–retest reliability. The test is thought of as parallel with itself. Some authors refer to the test–retest correlation obtained as the coefficient of stability.

With tests of achievement (e.g., a math test), aptitude (e.g., intelligence tests), and personality (e.g., a test of extroversion/introversion), the measure is typically administered only twice, so that only one estimate of the reliability coefficient is obtained. If the measure is administered several times, then the usual practice is to calculate the average or mean of the intercorrelations among the scores obtained on the various occasions as the estimate of the test–retest reliability coefficient.

## Assumptions and Considerations

The applicability of test–retest reliability depends on two primary assumptions. The first assumption is that the participants' true scores are stable across the two testing occasions. That is, the researcher must be confident that respondents' true scores do not change from the first measurement to the second measurement. The second assumption is that the error variance of the first test is equal to the error variance of the second test.

The assumption of stability of the construct that is measured is an important consideration in whether test–retest reliability is chosen as the method to assess the consistency of a measure. All changes from one testing period to another are treated as random errors. This type of reliability estimate assumes that the characteristic being measured is stable over time. Intelligence is one characteristic of a person assumed to be stable over time as is certain personality traits. Extroversion is one such personality trait as is the level of openness to new experiences. Other characters, however, are considered to be state, not trait, characteristics. That is, they are expected to fluctuate over time. Mood state is one such characteristic. Test–retest reliability is not an appropriate way to assess the reliability of a measure on mood state because any error variance would be caused by not only the unreliability of the measure but also the fact that mood is not a stable characteristic. For example, during the test–retest interval, one individual might have been physically ill and that influenced his mood. In contrast, another individual was given a major award during the same test–retest interval, which would inflate that person's mood level. As a result, individuals' mood states during the first assessment might be vastly different from the mood states at the second administration of the test. That is, the differences among their true construct levels are not stable across the two test administrations. For such state-like constructs, the test–retest method is a poor estimate of reliability. An internal consistency reliability method would be a better choice because the test might accurately reflect the difference in true score at that one occasion and therefore be reliable. However, the test–retest method might provide a low estimate of reliability because the moods have changed over the time interval between test administrations.

There are two major advantages of using the test–retest method of estimating reliability. The first is that only the test itself is required, unlike

other methods of estimating reliability that require more than one form (i.e., parallel or alternate forms). The second advantage is that the particular sample of items or stimulus situations is held constant, which would seem to minimize the possibility of measuring traits other than what is designed by this instrument.

One problem with estimating reliability using the test–retest method is deciding on the appropriate time interval between test administrations. If the interval is too short (e.g., a few days), then participants might remember their responses the first time the test is administered, which might influence their responses on the second administration. These effects are also known as carryover effects, which would likely inflate the reliability estimate. For example, a student taking a vocabulary test might have answered "true" to the item "deft means possessing skill or ease in performance" in the first test administration, and the second time he or she remembers this and answers "true" again. Another example is a supervisor in a retail store who assigned her supervisee John a rating of "above average" in June and assigns the same rating in January to demonstrate consistency in her evaluations. Of course, it is possible to have carryover effects that might increase the true scores over the time interval between administrations and, therefore, lower the test–retest reliability estimate. Sometimes, there are practice effects (also a type of carryover effect), so that on subsequent administrations of the same test, scores increase in a systematic fashion. The person might learn the specific content of the test or develop improved approaches to the material so that his or her score increases. These practice effects might be different for different individuals. Of two people with exactly the same score on one occasion, one person might think about the material between the administrations and find out the answers to certain questions. Therefore, on the second administration, one score might improve and the other will remain the same. So if the test–retest reliability coefficient is low, it is unknown whether the test is unreliable or differential systematic factors are at work.

One way to minimize carryover effects would be to increase the time interval between the two administrations of the test. Although this solves some problems, it creates others. Whereas it is true that the longer the time interval between administrations, the lower the probability of carryover effects such as the use of memory. However, it is also true that a longer time period might increase the likelihood that the participants will experience real change in the trait or characteristic being studied. Therefore, a low test–retest reliability correlation might be indicative of a measure with low reliability, of true changes in the persons being measured, or both. That is, in the test–retest method of estimating reliability, it is not possible to separate the reliability of measure from its stability. That is why it is usually suggested that the time between the two test administrations be relatively short (e.g., 1 or 2 weeks) so that it is more likely that only random error is being tapped and not true changes in the characteristic of interest. This is a serious disadvantage with the test–retest approach to estimating reliability. Therefore, this approach should be used with caution.

Test–retest reliability is an important consideration when a researcher is interested in measuring change over time. A simple measure of change is to have research subjects complete an instrument measuring the construct of interest at time 1. Then, an intervention to promote (hopefully) change is administered to the subjects. After the intervention, the instrument is given again. The difference between the two administrations of the test, which is often known as the gain score, is then taken as a measure of change. To be confident that the gain score reflects actual change in the participants for the construct of interest, the role of measurement error must be considered. That is, the researcher must be sure that the change is not simply a result of random fluctuation in an unreliable test. To assess the meaningfulness of the change score, the researcher needs to know the standard error of measure (SEM) of the test. The SEM must be calculated using the test–retest reliability coefficient rather than an internal consistency coefficient because it is the stability of the test that is most important here. The formula for arriving at the measure of change is widely referred to as the reliable change index (RCI), as described by Neil Jacobson, William Follette, and Dirk Revenstorf. RCI is calculated as follows:

$$Z = (x_2 - x_1)/(SEM\sqrt{2}),$$

where $Z$ is the RCI, $x_1$ is the time 1 administration of the test, and $x_2$ is the time 2 administration of

the test. RCI is a normally distributed index with a mean of 0 and a standard deviation of 1. If the absolute value of RCI exceeds 1.96, the observed change has a probability of .05 or less, and thus, the researcher can assume that the change in scores from time 1 to time 2 was not caused by random error. The higher the test–retest reliability coefficient, the more meaningful the gain score.

*Karen D. Multon*

***See also*** Coefficient Alpha; Correlation; Instrumentation; Reliability; Split-Half Reliability

**Further Readings**

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 105–146). Washington, DC: American Council on Education.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences.* San Francisco, CA: W. H. Freeman and Company.

Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1984). Toward a standard definition of clinically significant change. *Behavior Therapy, 17,* 308–311.

# THEORY

Theory is generally understood as a systematic representation of a genuine problem, articulated as far as possible in mathematical terms in the natural sciences or logical (or strictly linguistic terms) in the life and social sciences. The systematic nature of theory is normally aimed at providing explanatory leverage on a problem, describing innovative features of a phenomenon or providing predictive utility. The empirical adequacy required of a theory is a controversial feature of theories and often differs radically across disciplines. As most research in the sciences and social sciences is theory driven, that is, is concerned with the refinement or refutation of theoretical claims, the design of that research will have an immediate impact on the nature of theory construction and the presumed relationship among theory construction, observation, and the outcome of empirical research.

## Positivism and Its Impact on Theory

The traditions of research in the social sciences have consistently argued for a theory-driven approach to the design of investigations, particularly in its more experimental formulations. This tradition became established in the 20th century through the adoption of logical-positivist premises and a hypothetico-deductive framework in research that focused on observable features of the world as the final arbiter of research outcomes. Late 19th-century positivism, originally articulated by Auguste Comte, was influential in guiding the new social sciences, particularly psychology, in adopting research methods that would allow observed features of the world to determine the outcome of research questions. This was done to remove the influence of metaphysics and theology in science and originally had a salutary effect. However, this naïve version of positivism suffered from shortcomings that would make themselves felt in the 20th century. In particular, it ruled out the possibility of allowing entities such as atoms the status of real objects in science because they were unobservable. Ernst Mach, the Austrian physicist and philosopher whose philosophical views were important in the early social sciences, championed this version of positivism.

Positivism reduced the problem of theory to one of sensations, namely, the experience of an experimental outcome would in fact be the basis for a scientific theory. This was useful in removing the effects of metaphysics from science but hindered the progress of science as it came to rely ever more on unobservables. Logical positivism (in the version promulgated by the Vienna Circle) sought to ameliorate this. The principle of verification advanced by the movement claimed that the meaning of a statement is determined by sense-experience. However, this sense-experience can be direct or indirect allowing for the introduction of concepts (theory) that are linked to observations while not directly observable themselves. Instead, protocol sentences (direct observation terms) were linked to axioms (theoretical terms) with logico-mathematical terms. One consequence of this view is the rigid separation of theory from observation or data. Although it was dominant until after World War II, this position created an important sense that theory was to be understood as

a rigorous feature of experimental science. This was especially the case for theory that was formulated in mathematical terms or in terms of so-called laws. An additional concern that sprung out of the work of the logical positivists was a preoccupation with the unity of science. Dominant for much of the 20th century, the unity of science view was that all sciences are related through a thoroughgoing reductionism that would ultimately base all scientific theory on the laws of physics (although Otto Neurath, a logical positivist, was generally opposed to a crude reductionism of this sort).

Theory in the social sciences then was a highly restricted affair. Limited by its necessary ties to the nature of observations, restricted by logic, and pursued for its ends as a formalization of research, theory remained deeply tied to the immediate problems of the laboratory. For example, the learning theories in psychology that emanated from the behaviorist laboratories of Clark Hull, Edward C. Tolman, or Burrhus F. Skinner were, despite their elegance, often restatements or redescriptions of what the investigators thought the results of their experiments demonstrated. That is, the learning theories themselves were formulated in such a way as to ensure that the empirical demonstrations never deviated in large measure from the claims of the theory. Skinner famously called for an end to theories of learning altogether, although he continued to theorize apace.

Indeed, logical positivists themselves recognized that there was a problem with theory in their account. Carl Hempel's "paradox of theorizing" argues that if the terms and general principles of a scientific theory serve their purpose by establishing connections among observable phenomena, then they can be eliminated. This is because a chain of laws and interpretive statements that establishes such a connection should be replaceable by a statement that directly links observational antecedents to observational consequences. The dilemma then, for Hempel, was that if the terms and principles of a theory are unnecessary, then theory is unnecessary. Hempel claimed that perhaps theory is still useful because of its inductive, predictive, or heuristic uses, but this did not limit the damage to notions of logical positivism. Obviously, no one was about to abandon theorizing. Furthermore, others such as Karl Popper already noted that theories seemed more like "conjectures" than the formal arrangements of the logical positivists. Imre Lakatos argued famously that the core of scientific theories is generally protected by what he called auxiliary hypotheses and, hence, theories can never be just connecting statements between observational antecedents and consequences. By the 1970s, it was clear that a conception of theories in science was emerging that emphasized the surplus meaning of theory above and beyond the mere redescription of empirical content.

## End of the Received View

After World War II, severe doubts crept into the "received view," as the logical positivist version of science was often called. Norwood Hanson in his *Patterns of Discovery* (published in 1958) was among those who argued that the observations themselves are "theory laden." One does not naïvely observe the world as it is but always approaches the world with some preconceptions in place. Hanson argued that it is not a question of using or testing a theory but is the creation and discovery of what becomes an appropriate theory. In that sense, he claimed that he was more concerned with the way in which theoretical systems are built into observations such that they come to lead one to understand certain facts. Hanson did not claim that theories created facts de novo but that observations were always already colored by the theories or concepts that were used to frame the observations.

The question of theory and its relationship to data was revisited anew by Thomas Kuhn. In his influential book *The Structure of Scientific Revolutions* he increased the revisionist impulse in the understanding of science. Arguing that scientific advances run in cycles, Kuhn proposed that deeply embedded worldviews allowed science to proceed in a "normal" fashion until such time that the cracks in the established view led to a radical realignment, and a "scientific revolution" turns over the established "paradigm." Although Kuhn's work was later criticized, not least for its ambiguous use of the notion of paradigm, the overall impact was to steer philosophers and historians of science away from an emphasis on prescribing how science should proceed to describing how it

might actually work in the context of human institutional and social activities. Contemporary research in science studies has continued this tradition by emphasizing the social nature of scientific progress.

## Theories and Models

If theories are not precise statements of law, or systematic organizers of fact, then their role must be found elsewhere. Some philosophers of science, who were influenced by the work of Frederick Suppe, argued that theories were set-theoretic structures. This semantic conception of theories treats theories as abstract structures that stand in mapping relations to phenomena. In addition, on this view, scientists are busy constructing models that are idealized, partial simulations of the world or that probe real-world systems. According to Suppe, theories and models consist of mathematical spaces or structures standing in mapping relations to other systems. The consequence of this claim is that the structure of a scientific model can be examined as a semantic conception. These discussions have tended to focus in recent years on the problem of testing theories in such a way that the test is not tautological.

## The Ascendance of Functionalism

In the social sciences, however, where few models are expressed in precise mathematical terms, the relationship between theory and data or, more generally, theory and observation, continues to be a feature of ongoing debate. Theories come in different types and no one has determined the point at which the honorific label of theory is applied to a hunch or hypothesis. Traditions vary among social sciences, and what might be considered a theory might be anything from a counterintuitive proposal (for example, several features of evolutionary psychology) to a redescription of experimental outcomes. Certain neoclassical economic theories tend to favor computational solutions to problems. However, early in the 20th century, the social sciences created functionalism as a way of accounting for processes or properties whose activity was of interest even if their actual ontological status was uncertain. That is, even if one did not know precisely what a phenomenon consisted in, one might be able to explain how it functioned.

In recent decades, the predominance of functionalism in the social sciences has been firmly established. Functional accounts normally do not commit themselves to what there is but to what is occurring in the processes under investigation. For example, without even giving an account of just what memory *is,* theories of memory have long been expressed in terms of the functions of memory (short-term memory, long-term memory, episodic memory, semantic memory, etc.). Furthermore, the development of the cognitive sciences created another meaning of function and further cemented the tradition as one of the most dominant in the social sciences. For the cognitive scientist, functionalism was a thesis about the instantiation of a computer program on a physical platform, regardless of the nature of that platform. The mind could be likened to a computational function. A sense of this meaning of functionalism still permeates cognitive psychology and cognitive neuroscience.

## Willard Quine Empiricism

The use of the term *theory* then has become rather unrestricted. It is often a placeholder for a variety of unknowns that serves to keep us from leaping to firm conclusions while otherwise formalizing hunches and guesses. There might be some "empirical evidence" for the case at hand, but the theory supported by that evidence is, of course, always underdetermined. This means only that no empirical results bring finality to theoretical frames; this problem was originally articulated by philosopher Willard Quine in his well-known article "Two Dogmas of Empiricism" (the dogma of the distinction between analytic and synthetic truths and the dogma of reductionism). Although the debate that followed its publication was lengthy and detailed, the outcome was a gradual historical reevaluation of the sciences. Most important, however, was the support it would provide for the work of the historians and sociologists of science who demonstrated that the distinction between *theory* and *fact* is a rather dubious and unhelpful one in evaluating science, its results, its methods, and its products.

## Reflexivity

What characterizes most social science research is the reflexive nature of human activities. That is, to understand the human activities, the investigating scientist must of necessity be an apposite member of the community of participants under investigation. Otherwise, such phenomena as widely divergent as depression, abuse, socialization, fondness, racism, sexual attraction, and so on will be unintelligible to the investigator and, hence, will not likely capture all the relevant features of the phenomenon in question. One could argue that one needs to be a member of a community of scientists to do any kind of research; although the atomic number of, say, Fe (iron)—26—will not change according to cultural customs, depression might, in fact, do so. The latter is an example of a phenomenon that is the outcome of the shared practices of a particular community regardless of its neuroscientific precursors. Furthermore, as Joseph Margolis has argued, the combined infrapsychological powers alone of the aggregated members of a community cannot account for the features at hand; it is not just a process that occurs "inside the head." Instead, the phenomenon is the outcome of a communal set of activities; to be depressed is by definition to be depressed in a particular culture.

Other philosophers of the social sciences, such as Rom Harré, have argued that what characterizes the social sciences is a concern with the activities of ordinary life in their most varied and rich circumstances. These activities are best captured by the Wittgensteinian distinction of "following a rule" versus "acting in accord with a rule." The former is done when conforming to a set of prescribed rules, such as those that must be followed when learning some new, complex skill. The latter is done when one speaks a natural language as a child, and thus, one has never observed the grammatical rules that are nevertheless followed, or when driving a car effortlessly after many years of practice even though one learned to do so by initially following a concrete set of rules. The distinction here is between habitual activity, which is highly overlearned, and the kind of activity that involves conforming to norms and conventions one might barely be able to articulate. A theory that takes the latter into account must have features unlike a neuroscientific account of human activity because norms and conventions are by definition communal pheonomena.

In psychology, however, there are many who do not believe that special theoretical concessions need be made for these kinds of distinctions and that eventually all psychological phenomenon will be captured by some overarching neuroscientific or evolutionary framework. Despite the popularity of this reductive claim, these theories are a considerable explanatory distance from the activities of ordinary life. For an account of these kinds of activities, goes the counterargument, attention needs to be paid to the microdetails of life itself either through talk (as in various discourse analytic positions) or other forms of ethnography. Proponents of overarching theories, in contrast, argue that their positions will provide an ultimate explanation of the foundations of such activities.

## Broadened Conceptions of Theory

Events outside the social sciences have also led to new and unusual forms of theory in recent years, such that the term has come to be used in an ever wider sense. Under the influence of formalism, historicism, Marxism, structuralism, feminism, queer theory, ethnic studies, postcolonialism, culture studies, and more, the idea of a theory came to represent something of a challenge to orthodoxy. In part motivated by the crisis in epistemology that was a consequence of postmodernism and the denial of the possibility of certain knowledge, theory in the humanities resisted so-called universalizing impulses. Critical theory has been inserted into the humanities since the 1960s in a way that has led to new emphases on theoretical scholarship. Having suffered through a period of excess in the late 20th century, some of these orientations have settled into academic subdisciplines, whereas others have passed altogether. One important consequence has been the general suspicion of grand theories espoused traditionally in the social sciences. Hence, from Sigmund Freud to Talcott Parsons, the overarching theoretical frame has been treated with wariness and viewed as carrying within it other interests (for example, the phallocentric nature of Freud's theory) that are not explicitly accounted for in the theory itself.

The consequences of these developments for the social sciences have been less disruptive to the

research enterprise than they have been for the humanities. Nonetheless, the effects can be observed in the further institutionalization of theoretical traditions in such disciplines as sociology and psychology, where new divisions and societies devoted to the study of theory have come into their own in recent decades (e.g., the International Society for Theoretical Psychology) as well as new journals (e.g., *Theory & Psychology*). In addition, the greater influence of qualitative methods or microsociological research has changed the way theory is organized in the academy. The question of theory is not so much a question of finding the best "fit" for the data as it is providing an adequate account of a phenomenon that reflexively takes up the position of the researcher as well as the participants into its final articulation.

Theory is no longer the prerogative of narrowly determined empirical research enterprises that extract hypotheses from an overarching theory and then carefully test these along a neo-positivist model. In all areas of the social sciences, theory includes models and hypotheses, as well as systematically developed frameworks or just initial articulations of what is the phenomenon of interest. What constitutes theory is now largely a matter of convention as well as locally or disciplinarily determined rules for empirical adequacy.

*Henderikus J. Stam*

***See also*** Cause-and-Effect; Content Validity; External Validity; Grounded Theory; Models; Observations; Positivism; Thought Experiments; Validity of Research Conclusions

### Further Readings

Bem, S., & Looren de Jong, H. (2005). *Theoretical issues in psychology: An introduction* (2nd ed.). Thousand Oaks, CA: Sage.

Hempel, C. G. (1965). *Aspects of scientific explanation and other essays in the philosophy of science*. New York: Free Press.

Quine, W. V. O. (1980). Two dogmas of empiricism. In W. V. O. Quine (Ed.), *From a logical point of view* (2nd rev. ed.) (pp. 20–46). Cambridge, MA: Harvard University Press.

Suppe, F. (1998). Theories, scientific. In E. Craig (Ed.), *Routledge Encyclopedia of Philosophy*. London: Routledge. Retrieved from http://www.rep.routledge.com/article/Q104SECT7

# Theory of Attitude Measurement

Attitude can be defined as a tendency to react favorably, neutrally, or unfavorably toward a particular class of stimuli, such as a custom, institutional practice, or national group. There are two challenges a researcher faces when measuring an individual's attitude. First, an individual's attitude toward an object cannot be observed directly but must be inferred from observed behavior, such as responses to a questionnaire. And second, there is no inherent scale associated with the observed behavior. The goal of a scaling method is to define a continuum/scale in which to place stimuli and respondents. The purpose of this entry is to describe the Thurstonian scaling method of similar reactions/attributes described by Louis Leon Thurstone, in which the scale values for each statement represent the actual distance or separation between each statement with respect to the degree of favorableness toward the object.

## Method of Similar Attributes/Reactions

Suppose a researcher has $I$ statements regarding the attitude toward a particular object (e.g., capital punishment). The first step in the method of similar attributes is to rank order the statements from least favorable to most favorable regarding the object of interest. The method of equal-appearing intervals (also described by Thurstone) might be used to rank order the statements. Next, the coefficient of similarity (denoted $\phi$) is computed between all pairs of statements using raw data collected from a sample of participants who were asked to endorse statements he or she agrees with. The $\phi$-coefficient represents the degree to which the endorsement of two statements reflects the same attitude. For any statement $i$ and statement $j$, the $\phi$-coefficient of similarity is computed as follows:

$$\phi_{ij} = \frac{n_{ij}}{\sqrt{p_i p_j n_i n_j}}, \quad (1)$$

where $n_{ij}$ represents the number of participants who endorsed both statement $i$ and $j$; $n_i$ and $n_j$ represent the number of participants who endorsed

statement $i$ and statement $j$, respectively; and $p_i$ and $p_j$ represent the reliability of statement $i$ and statement $j$, respectively. The reliability of a statement represents the probability a participant will endorse the particular statement given he or she has that opinion. In computing the reliability, the fact that the statements have been rank ordered such that any two adjacent statements will reflect practically the same attitude is exploited. Therefore, the reliability of statement $i$, where statement $j$ is adjacent to statement $i$, is computed as follows:

$$p_i = \frac{n_{ij}}{n_j}. \tag{2}$$

Essentially, the reliability of an item represents the proportion of participants who endorsed an adjacent statement that represents the same attitude.

To illustrate the procedure, Table 1 reports hypothetical raw data corresponding to five statements that were selected from a scale consisting of 50 statements administered to 1,000 participants (note that five statements have been selected for illustrative purposes only—i.e., a scale should contain many more items). The diagonal elements in Table 1, which are bolded, report the number of participants who endorsed each statement. The off-diagonal elements indicate the number of participants who endorsed both respective statements. The reliability of each statement is reported in the last row labeled $p_i$ and is based on Equation 2, which was applied to the complete scale of 50 statements. The statements in Table 1 have been ordered from least favorable ($s_1$) to most favorable ($s_5$) based on the method of equal-appearing intervals, which was applied to all 50 statements.

**Table 1** Number of Participants Who Endorsed a Pair of Statements

| | *Statement* | | | | |
|---|---|---|---|---|---|
| | $s_1$ | $s_2$ | $s_3$ | $s_4$ | $s_5$ |
| $s_1$ | **221** | 120 | 90 | 44 | 15 |
| $s_2$ | 120 | **294** | 214 | 146 | 55 |
| $s_3$ | 90 | 214 | **350** | 215 | 101 |
| $s_4$ | 44 | 146 | 215 | **423** | 274 |
| $s_5$ | 15 | 55 | 101 | 274 | **515** |
| $p_i$ | 0.80 | 0.73 | 0.64 | 0.81 | 0.43 |

**Table 2** $\phi$-Coefficient Indicating Similarity for All Pairs of Statements

| | *Statement* | | | | |
|---|---|---|---|---|---|
| | $s_1$ | $s_2$ | $s_3$ | $s_4$ | $s_5$ |
| $s_1$ | 1.00 | 0.62 | 0.45 | 0.18 | 0.08 |
| $s_2$ | 0.62 | 1.00 | 0.98 | 0.54 | 0.25 |
| $s_3$ | 0.45 | 0.98 | 1.00 | 0.78 | 0.45 |
| $s_4$ | 0.18 | 0.54 | 0.78 | 1.00 | 0.99 |
| $s_5$ | 0.08 | 0.25 | 0.45 | 0.99 | 1.00 |

Table 2 reports the coefficient of similarity ($\phi$) based on the data from Table 1. The $\phi$-coefficient represents the degree to which any two statements are similar. The $\phi$-coefficient typically ranges from 0.00 to 1.00, where relatively large values of $\phi$ indicate that the two statements are similar and thus are close together on the continuum; small values of $\phi$ indicate that the two statements are dissimilar and thus are far from each other on the continuum. It is assumed that the $\phi$-coefficient has a maximum of 1.00 when the scale separation is zero (e.g., when a statement is compared with itself or is the same as another statement).

The $\phi$-coefficients are converted to scale values to measure the separation between any pair of statements. The scale separation between any pair of statements can be obtained by finding the standard normal score associated with the ordinate (i.e., height of the standard normal distribution) defined by the $\phi$-coefficient. Thurstone suggested using a probability table in which the maximum ordinate is 1.00 (however, the maximum ordinate provided by most software packages is .40—dividing the ordinates by .40 will convert them such that the maximum equals 1.00). Table 3 reports the scale separation values for all pairs of statements.

When the $\phi$-coefficient is near 1.00, the scale separation will be close to zero, which indicates that the two statements are similar, whereas for small values of $\phi$ the scale separation will be large. For example, for $\phi = 0.98$ (e.g., statements 2 and 3), the scale separation ($S_3 - S_2$) is 0.19, whereas for $\phi = 0.54$ (statements 2 and 4), the scale separation ($S_4 - S_2$) is 1.11. The sign of the scale separation value is determined by the end of the scale that is arbitrarily defined to be positive. For example, if the statement that has been rank ordered

**Table 3** Separation Values ($S_{column} - S_{row}$) for All Pairs of Statements

| | *Statement* | | | | |
|---|---|---|---|---|---|
| | $s_1$ | $s_2$ | $s_3$ | $s_4$ | $s_5$ |
| $s_1$ | 0.00 | 0.97 | 1.26 | 1.85 | 2.24 |
| $s_2$ | −0.97 | 0.00 | 0.19 | 1.11 | 1.66 |
| $s_3$ | −1.26 | −0.19 | 0.00 | 0.70 | 1.26 |
| $s_4$ | −1.85 | −1.11 | −0.70 | 0.00 | 0.12 |
| $s_5$ | −2.24 | −1.66 | −1.26 | −0.12 | 0.00 |

first, which indicates that it represents the least favorable regarding the object, is defined as zero, then the scale separation values associated with this statement will be negative (indicating that the scale value is less than all of the other statements). Last, Thurstone recommended ignoring scale separation values greater than or equal to 2 because of unreliability.

The final scale value for each statement is based on the average scale separations between adjacent statements defined by the rank ordering. The average scale separation for adjacent statements $i$ and $j$ might be expressed as follows:

$$\text{Ave}(S_i - S_j) = \frac{1}{I}\left[\sum_{k=1}^{I}(S_i - S_k) - \sum_{k=1}^{I}(S_j - S_k)\right], \quad (3)$$

where $(S_i - S_k)$ represents the scale separation values for statement $i$ and $k$ reported in the body of Table 3. For example, the average scale separation for statements 3 and 4 is

$$\text{Ave}(S_4 - S_3) = \frac{1}{5}[(3.54) - (-0.51)] = 0.81. \quad (4)$$

Table 4 reports the final scale values for the five statements. The final scale value is determined by first arbitrarily setting the scale value for statement 1 (i.e., the statement that represents the least favorable attitude) to be zero and then aggregating the average scale separation values across the statements. For example, the final scale value for statement 3 equals the statement 2 final value plus the average scale separation between statements 3 and 2 (i.e., 1.234 = 0.938 + 0.296).

The final scale values define the continuum and allow researchers to determine where a participant scores in comparison to the stimuli.

**Table 4** Final Scale Values for the Five Statements

| *Statement* | *Average Scale Separation* | *Final Scale Value* |
|---|---|---|
| $s_1$ | 0.000 | 0.000 |
| $s_2$ | 0.938 | 0.938 |
| $s_3$ | 0.296 | 1.234 |
| $s_4$ | 0.810 | 2.044 |
| $s_5$ | 0.338 | 2.381 |

*Craig Stephen Wells*

***See also*** Guttman Scaling; Interval Scale; Levels of Measurement; Likert Scaling; Normal Distribution; Thurstone Scaling

**Further Readings**

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology, 33,* 529–554.
Thurstone, L. L. (1929). Fechner's law and the method of equal-appearing intervals. *Journal of Experimental Psychology, 12,* 214–224.
Thurstone, L. L. (1929). The theory of attitude measurement. *Psychological Review, 36,* 222–241.

# Think-Aloud Methods

Think-aloud methods ask participants to verbalize their thoughts while performing a task. Such methods provide a basis for investigating the mental processes underlying complex task performance and can provide rich data on such cognitive processes. Since the inception of scientific psychology, think-aloud methods have contributed substantially to the understanding of problem solving and learning. This entry first describes the history of think-aloud approaches and then discusses the protocol analysis. Last, this entry addresses some limitations of think-aloud methods.

## History

Early pioneers of scientific psychology, such as William James, Wilheim Wundt, Alfred Binet, and Edward Titchener, used introspective reports of subjective experiences to provide insight into human consciousness, learning, and problem solving. However, the lack of reproducibility of findings from the analysis of introspective reports

resulted in its marginalization and scientific psychology's emphasis turned from consciousness to observable behavior.

However, it should be noted that James Watson, whose seminal paper "Psychology as the Behaviorist Views It" provided the impetus to remove introspection from scientific study, advocated the use of think-aloud methods to understand cognitive processes. Gestalt psychologists also used think-aloud approaches to better understand cognitive processing during problem solving. Furthermore, Jean Piaget's seminal studies of the development of cognitive abilities relied on children's verbal descriptions of their responses to various experimental tasks. Contrary to the analytic introspection research that required participants be trained in self-observation or provide immediate retrospective reports of thoughts, participants were now being asked to simply "think aloud" their thoughts *as they emerged* while performing a task.

Psychology's embrace of information-processing approaches to understanding behavior in the 1950s onward resulted in the re-emergence of think-aloud methods as a valued research tool. The method gained popularity through Allen Newell and Herbert A Simon's use of think-aloud data to inform their work on computational processes in problem solving. Contemporary use of the think-aloud method is informed by K Anders Ericsson and Simon's protocol analysis of verbal reports.

## Protocol Analysis

In a think-aloud study, participants are asked to verbalize thoughts that emerge as a task is being completed. The method aims to elicit the information required for task performance and consequently, the verbalizations should reflect the thoughts being attended to at the time. Participants should not provide an explanation of such thoughts, because to do so, they have to draw on additional thoughts and explanations that are not related to the task at hand and might alter the structure of the thought processes under study. Successful elicitation of the focal thoughts requires participants to attend uninterrupted to the completion of the task presented.

As people often feel that they have to explain or justify their thoughts in a social interaction, many strategies are used to minimize such possibilities. Participants often complete simple warm-up tasks to practice attending completely to a given task while verbalizing their thoughts only. It is also recommended that the researcher sits behind the participant to remove the social interactive element of the procedure. The researcher monitors the verbalizations and reminds the participant to speak if there has been a period of silence.

### *Concurrent Versus Retrospective Verbal Reports*

A distinction is made between concurrent verbalization (thoughts are verbalized that emerge as the task is being completed) and retrospective verbalization (after completion of the task, the participant is asked about the thoughts that occurred at an earlier point in time). Retrospective verbalization aims to avoid the potential problem of task performance disrupting thoughts, but it is problematic as people might not accurately remember the thoughts that arose in completing the task. Post hoc rationalization of the cognitive processes might occur unintentionally. Hence, concurrent verbal reports are believed to be more accurate, and where appropriate, both concurrent and retrospective reports should be collected.

### *Analysis*

Protocol analysis requires the verbal data to be classified into meaningful categories. The verbal responses are recorded verbatim, typically using a tape recorder, and transcribed. The data are reliably coded (e.g., using multiple raters and assessing interrater reliability of coding) into meaningful categories; such content analysis might be theory-driven or empirically driven depending on the nature of the study. For example, the think-aloud method can be used to develop a model of the cognitive processes underpinning problem solving or to test the validity of model derived from psychological theory. The verbalizations are often disjointed without any obvious relation between thoughts, and such incoherency and incompleteness is inherent in think-aloud data. The verbal protocols can be subjected to inferential categorical

analysis (e.g., chi-square test) or simply descriptive analysis (e.g., frequencies).

### *Applications*

Protocol analysis has been widely used to examine thinking in basic cognitive science and in many applied areas. Think-aloud methods have elucidated expert versus novice task performance across a variety of domains (e.g., chess). For example, in addition to having more knowledge, verbal protocol analysis has revealed that experts are better able to retrieve relevant information. In addition, protocol analysis has been applied to diverse topics, including cognitive processes in reading and writing, completing IQ tests, person perception, and scale development.

### Limitations

A frequent criticism of the think-aloud method is that the act of concurrent verbalization might change task performance and the cognitive processes studied. However, Ericsson and Simon argue that concurrent verbalizations do not change the course or structure of thought processes. As the method relies on the verbalization of thoughts accessible in working memory, automatic processes cannot be accessed. Participants might only verbalize some working memory information used in task completion and hence provide an incomplete representation of the salient thoughts. Furthermore, the demand characteristics (e.g., social desirability) of the research might influence the nature of verbalizations made.

*David Hevey*

***See also*** Chi-Square Test; Content Analysis; Frequency Distribution; Interrater Reliability

#### Further Readings

Crutcher, R. J. (1994). Telling what we know: The use of verbal report methodologies in psychological research. *Psychological Science, 5,* 241–244.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge: MIT Press.

## Thought Experiments

A thought experiment is an experiment that aims to persuade by reflection on its design rather than by its execution. After reviewing the origin of thought experiments and the empiricist-rationalist debate over their effectiveness, this entry concentrates on Ernst Mach's influential account.

### History

Actual thought experiments predate the term *thought experiment,* which was first used by Georg Christoph Lichtenberg in 1793 and only acquired wide currency after Mach's 1897 essay "Gedanken experiment." Two thousand years before the term was invented, poets imported thought experiments into metaphysical verses about the nature of the universe. In his epic De Rerum Natura, the Roman poet Titus Lucretius supports the infinity of space by having the reader imagine that space has a boundary. If a man threw a spear at the edge of space, either the spear passes through, and so there is no edge of space, or it bounces back, in which case there is a barrier that lies beyond space and yet is in space.

Rationalists welcome thought experiments. This a priori method shows that we can learn about the world without experience! Explaining *how* this happens is a classic problem—one that Plato hoped to solve with his Theory of Forms. Contemporary philosophers, such as James Robert Brown, continue to work on an explanation within the Platonic framework.

### Empiricist Criticism

Empiricists have trouble understanding how thought experiments could be any more informative than poetry. If everything is learned from experience, then no information about the world can be extracted from a *hypothetical* experiment.

Yet many attacks on Aristotle's theory of motion are launched from Galileo's armchair. Consider his critique of Aristotle's principle that heavier objects fall faster than lighter objects. Galileo supposes that a smaller object becomes attached to a heavier object during its descent. On

the one hand, the composite object must slow down because the heavy component is retarded by the slower component. On the other hand, the composite object must accelerate because it is heavier than the heavy component.

Empiricist reactions to thought experiments fall along a spectrum. Pierre Duhem condemns thought experiments as illusory sources of support. They have no more place in science than Plato's numerological arguments for the existence of seven planets.

John Norton thinks the success of Albert Einstein's thought experiments shows that Duhem was too dismissive. According to Norton, thought experiments should be heeded exactly to the degree that they can be reduced to cogent arguments. After all, empiricists endorse changes of opinion that are based on surprising deductions and calculations (even if the premises and data were already known).

## Mach's Empiricist Account of Thought Experiment

Einstein's early philosophical mentor, Mach, was more concessive than Norton. Mach believed that thought experiments tap into stored "instinctive knowledge."

The pictorial form of this knowledge is evident from how questions such as "How many corners are in the capital letter preceding G?" are answered. A mental image of the letter is formed and the corners counted. In a sense, the number of corners was not known (they had to be counted), and in a sense, it was known (all the information was available to quickly answer the question). Information is stored in mental images in the way information is stored in crime scene photographs (which, if need be, can be consulted later for missed clues).

The second form of instinctive knowledge is inferential. Thanks to personal experience, mental associations mimic natural associations. But these echoes extend well beyond personal experience. Thanks to the generate-and-eliminate mechanism of natural selection, inferential dispositions reliably reflect information that stretches back across thousands of generations. For instance, when babies begin to walk, they also begin to associate darkness with danger. They fear the dark regardless of whether they suffer or witness mishaps. The fear intensifies as their mobility increases and only begins to wane when they become educable. The child's fear is a form of innate knowledge—best explained by Charles Darwin's biology, not Plato's metaphysics.

Each human being is a sample of the universe. Evolution favors the more representative microcosmoses, so wider patterns of nature resonate in everyone. Thought experiments amplify the signal (by suppressing noise through negative idealizations, such as stipulating away friction, and by positive idealizations, such as caricaturing magnitudes to make effects salient). For instance, consider someone with a round trip ticket from Los Angeles to New York. He knows that prevailing winds accelerate travel west to east but retard travel east to west. To determine whether the losses and gains cancel out, he can realistically assume the jet speed at 500 miles per hour and exaggerate the wind speed to 499 miles per hour. The Los Angeles to New York flight is fast, but the return flight takes more than a month. The surreal thought experiment helps the traveler answer his mundane question by underscoring the principle that opposite forces cancel out only when applied equally.

## The Economy of Thought

Actual trial and error is slow, costly, and dangerous. Those who could test hypotheses "by thought" lived longer and more economically. Although these mental simulations are less reliable than actual experiments, they are reliable enough to have a legitimate role in science. Techniques that trade accuracy for speed are just as much a part of the scientific repertoire as randomized control trials. (This explains why experiments that are designed to expose weaknesses of thought experiments are commonly designed with the help of thought experiments.)

Planning an experiment is itself a process that leads to rational revisions of opinion. Once these accidental improvements in cognitive credentials are noticed, one tries to achieve them deliberately. That is, experiments are crafted to rationally persuade just by reflection on their designs, rather than by execution. For instance, after Galileo established the law of equal heights (that a ball will roll up a U-shaped track until it recovers its

original height), he asked what would happen if the ball were rolled along an infinite, frictionless plane in a vacuum. The answer suggested by thought experiment, that the ball would roll endlessly in a straight line, is an anticipation of Newton's first law of motion.

Galileo did not fully believe the outcome of his thought experiment. He continued to subscribe to Aristotle's idea that natural motion is circular. But because a straight line approximates the segment of a large circle (as large as the earth), Galileo regarded the thought experiment as giving a result that was close to the truth.

This discrepancy between the thought experimenter's expectation and the result eases some concern about circularity. If imagination perfectly reflected prior theoretical commitments, then thought experiments could not be used to *test* theories.

As in the case of executed experiments, thought experiments are often marred by biases and methodological shortcomings. Measures are taken to damp down these disturbances.

Mach's naturalistic account of thought experiments suggests that thought experiments should only be reliable in conditions similar to those in which intuitions evolved and developed. Thought experiments that rely on intuitions about how objects behave under the extremes of temperature, pressure, and gravity should be regarded with suspicion.

Yet many of Einstein's thought experiments do concern the behavior of objects that are moving at unprecedented speeds. There are other well-regarded thought experiments concerning phenomena to which hunter-gatherers were oblivious (ones involving microscopic scale, electricity, quantum phenomena). If these thought experiments are as successful as many physicists assume, then they are anomalies for Mach's account.

The same goes for thought experiments in ethics and aesthetics. Nature is indifferent to what *should* be the case and what *should* be admired.

These challenges are akin to those first posed to Darwin. Thought experiments will themselves play some role in answering these questions. Like other scientific methods, thought experiments must pull themselves up by their own bootstraps.

*Roy Sorensen*

***See also*** Inference: Deductive and Inductive; Scientific Method; Theory

**Further Readings**

Brown, J. R. (1991). *Laboratory of the mind*. London: Routledge.

Mach, E. (1976). On thought experiments. In *Knowledge and error* (J. McCormack, Trans.) (pp. 134–147). Dordrecht, the Netherlands: Reidel.

Norton, J. (2004). On thought experiments: Is there more to the argument? Proceedings of the 2002 biennial meeting of the Philosophy of Science Association. *Philosophy of Science, 71,* 1139–1151.

Sorensen, R. (1992). *Thought experiments*. Oxford, UK: Oxford University Press.

# Threats to Validity

From a research design standpoint, the simplest way to understand threats to validity is that a hypothesis might be tested in a manner other than what the researcher had intended—a situation not to be confused with the researcher's failure to obtain the result he or she had expected. Much is presupposed in this distinction. Research is not a linear extension of common sense or everyday observation but rather requires a prior theory or paradigm that yields an appropriate hypothesis, on the basis of which the researcher selects relevant variables that are then operationalized and manipulated in an environment of the researcher's creation and control. Under these circumstances, two matters of validity arise: first, the reliability of the outcomes vis-à-vis other experiments of similar design, and second, the generalizability of the outcomes to the population that the experiment purports to model. The statistician Ronald Fisher established this way of thinking about research design validity, which included the random assignment of subjects to groups of the researcher's choosing, given the variables that need to be operationalized to test the researcher's hypothesis.

In what follows, the peculiarities of Fisher's original approach are examined, followed by the philosophical foundations of the concept of validity that is presupposed in research design, culminating in a discussion of the different threats to validity associated with different research designs.

## Peculiarities of Fisher's Original Approach to Research Design

Two features of Fisher's original context help to explain the initial clarity but long-term difficulties of his approach. These might be called "meta-threats" to validity, because they pertain to the selection of the paradigm of research design validity. First, Fisher developed his paradigm in the context of an agricultural research station, and hence, the things subject to "random assignment" were seeds or soils, not human beings. Second, the station was (U.K.) state run and not for profit. Whereas the former pertains to the concerns originally raised by Donald Campbell and Julian Stanley and in the bulk of this entry, the latter point has received increasing attention—indeed, as part of a general challenge to the idea that statistical significance is a meaningful outcome of hypothesis testing. This latter difficulty is briefly discussed in the next paragraph, before resuming with the threats to research validity specifically related to human subjects.

In the context where Fisher designed experiments on the efficacy of genetic and environmental factors on agricultural output, matters of utility and cost did not figure prominently. A state interested in acquiring a comprehensive understanding of what is likely to make a difference to food production presumed that every hypothesis was equally worthy of study. Thus, "statistical significance" came to be defined as the likelihood of an experimental outcome, given a particular hypothesis, which is tested by seeing whether the outcome would have been the same even if the hypothesis were false (i.e., the null hypothesis). In contrast, someone more explicitly concerned with utility and cost might cast the idea of validity in hypothesis testing as a species of normal rational decision making. In that case, the validity would be defined as follows: Given the available evidence, the cost of accepting a hypothesis vis-à-vis potentially better hypotheses that might require additional testing. Indeed, as Stephen T. Ziliak and Deirdre N. McCloskey observe, this stance was taken by William Sealey Gosset, a student of Karl Pearson who was lab director at Guinness breweries in the early 20th century.

## Philosophical Foundations of the Concept of Validity in Research Design

*Validity* is a curious English word to use for, broadly speaking, knowledge acquired by methodologically appropriate means. Etymologically, the word is related to *value,* not *truth*. In formal logic, a valid inference is one that preserves the truth of the premises in the conclusion: Truth matters to validity not in itself but only insofar as it is preserved from one logical moment to the next. From a false premise, any conclusion—true or false—validly follows. Validity merely captures the fact that content has not been lost in the translation from premise to conclusion. In the case of a false statement, because nothing true has been asserted, there is nothing to lose in translation. Thus, anything goes.

Although the logician's way of understanding validity might make good sense when considering bare knowledge claims, more complex and concrete acts of research like experiments, ethnographies, simulations, and surveys demand a more robust sense of "normative force," or what the Germans call *Geltung*. Etymologically, the model for *Geltung* is the state's backing of banknotes that enables them to retain their face value as they change hands in transactions over a wide variety of goods and services. There is a promise to pay a fixed amount, should one decide to cash in one's banknotes. By analogy, one could ask how a piece of research establishes its significance beyond the original research moment, so that it becomes a fixed contribution to knowledge, on the basis of which subsequent claims to knowledge can be evaluated and built. Campbell and Stanley would put the matter in terms of the grounds on which research can be used as a basis for policy. The starting point here is the recognition that any experiment or even ethnography, simply taken alone, is a statistically improbable act in the normal run of human events. This, then, raises the question of what, if any, standing it should have in explaining or directing human affairs.

Questions about the determination of research validity began with debates over the nature of *Geltung* roughly a century ago, relating to the Neo-Kantian view that the methodology of science is the main substantive project of epistemology. A brief recapitulation of the original philosophical

context puts the matter in focus. Immanuel Kant's theory of normative force (i.e., his general theory of what should be the case) presumed that David Hume was right to argue that "is does not imply ought": The bare fact that something exists has no normative force *per se*. Normative force depends on whatever else it manages to empower. As Kant put it: "Ought implies can." This point was driven home in several disciplines, most notably jurisprudence, where it informed the positivist model of a legal system as constituted by a sovereign whose will underwrites all legislation. Accordingly, the law's legitimacy rests on both the means by which the sovereign will is enforced, specifically the sanctions entailed by nonconformity, and the basis on which one comes to believe in the efficacy of those means. Science differs from law in that the difference between these two sources of legitimacy is minimized as the scientist's research design allows her first-hand knowledge of how the relevant claims acquire the normative force they have.

Neo-Kantian philosopher Heinrich Rickert, perhaps the most influential epistemologist prior to the rise of logical positivism, generalized this point to argue that the sciences (i.e., *Wissenschaften*) might be distinguished, not by the nature of their objects, but by the nature of their validity conditions. Thus, a study of human beings might be "valid" (or not) in many different senses depending on the intensity of the normative force attached to the study itself and the extent of the study's normative reach. Moreover, these two dimensions might trade off against each other, which is inscribed in the Neo-Kantian literature as the distinction between *idiographic* and *nomothetic* inquiry. Someone immersed in a particular native or historic culture (i.e., idiographic) more reliably knows that culture than someone who regards it as simply one of a kind (i.e., nomothetic), yet that superior knowledge is unlikely to be generalizable to other cultures.

The clearest contemporary legacy of this perspective is Jürgen Habermas's four validity claims (*Geltungsanspruchen*) relevant to human action, which imply that research might be (in)valid in terms of the accuracy, appropriateness, sincerity, and/or comprehensiveness vis-à-vis its target reality. According to Habermas, the existence of relatively self-contained yet equally enlightening schools of sociology, each dedicated to establishing one of these validity claims, demonstrates that validity is not a univocal concept and that so-called threats to it might be best understood as an attempt to mix different validity claims in a single study. This point is implicitly recognized at the most fundamental level of research design, in which a study might provide an adequate model of either other studies of exactly the same kind or of a range of rather different studies performed in conditions more closely resembling the target reality one wishes to capture or affect. This is the basis of the distinction between "internal" and "external" validity. Valid research designs in the former sense are often called "reliable," in the latter sense "representative," but there is no general expectation that these two qualities will be positively correlated.

## Types of Validity Threat in Research Design

The actual phrase *threats to validity* is a somewhat alarmist Cold War coinage from Campbell and Stanley. Campbell and Stanley had introduced the quasi-experimental design in response to demoralized education researchers who despaired of ever identifying a "crucial experiment" that could decide between rival hypotheses, each of which could claim some empirical support in explaining a common phenomenon. Campbell and Stanley concluded that such despair reflected an oversimplified view of the research situation, which demanded a more complex understanding of research design that enabled rivals to demonstrate the different senses and degrees in which all of their claims might be true. But opening up the research situation in this fashion legitimized the loosening of the various controls associated with the experimental method, which in turn reinvited versions of the traditional problems of naturalistic observation. It was these problems that Campbell and Stanley attempted to address as threats to validity.

A survey of the 12 threats to validity enumerated in Campbell and Stanley reveals that most of them involve insensitivity to the complex sociology of the research situation. For example, research might be invalidated because the researcher fails to take into account the interaction effects between researcher and subject, the subject's own response to the research situation over time, the subject's

belonging to categories that remain formally unacknowledged in the research situation but might be relevant to the research outcome, salient differences between sets of old and new subjects, as well as old and new research situations, when attempting to reproduce an outcome, and so on.

The quasi-experimental research design that Campbell and Stanley advocated as a strategy to avoid, mitigate, and/or compensate for such potential invalidations can be understood as a kind of "forensic sociology of science" that implicitly concedes that social life is not normally organized—either in terms of the constitution of individuals or the structures governing their interaction—to facilitate generalizable research. In effect, methodologically sound social research involves an uphill struggle against the society in which it is located and which it might subsequently improve with valid interventions. This rather heroic albeit influential premise has come under increasing pressure as insights from the sociology of science have been more explicitly brought to bear on research design. When presented as a reflexive application of the scientific method to itself, the resulting analysis can be skeptical.

Whatever practical problems are posed by the achievement of internal validity, the main theoretical problems are posed by the achievement of external validity. Here, it is useful to distinguish *ecological* and *functional* validity, the latter being external validity in the strict sense. This distinction, which was inspired by Campbell's teacher Egon Brunswik, pertains to the respects in which one would expect or wish the target reality to resemble the original research event, typically an experiment. Ecological validity pertains to an interest in reproducing the causes and functional validity the effects. Few experiments meet the standard of ecological validity but many might achieve functional validity if they can simulate in the target environment the combination of factors that had produced the outcome in the original experiment. The key epistemological point made by dividing external validity in this fashion is that experiments might provide valid guides to policy intervention without necessarily capturing anything historically valid about the underlying causal relations.

This point bears on Augustine Brannigan's skepticism concerning Stanley Milgram– and Phillip Zimbardo–style experiments that demonstrate the relative ease with which subjects will submit to authority and torture their fellows. Brannigan doubts that they provide the basis for inferring some deep, perhaps genetically based moral deficiency in *Homo sapiens*. Although his skepticism is probably warranted with respect to any questions of evolutionary psychology, it is equally probable that those experiments could provide—if they have not already—the basis for re-creating the outcomes in such nonresearch settings as the gathering of intelligence from potential enemies. This would provide a vivid example of an experiment manifesting functional validity yet lacking ecological validity.

This divided verdict on the external validity of experiments is also familiar from economics, especially thought experiments about what the ideal rational agent *Homo oeconomicus* would do under various hypothetical conditions. On this basis, Karl Menger, who brought the marginalist revolution in economics to the German-speaking world, sharply distinguished between the rationality of *H. oeconomicus* and the imperfect rationality of real economic agents. The phrases *normative* and *positive* economics were invoked to stress the gulf between these two states. Clearly, *H. oeconomicus* was not an ecologically valid construct. But additionally, and more importantly from a policy standpoint, Menger held that the interconnectedness of markets and the unique array of forces influencing each agent's decision making meant that even claims to functional validity could never be achieved, certainly not at the macroeconomic level. Such skepticism underwrites the continued aversion of the libertarian Austrian school of economics to the use of statistics as a basis for major policy interventions, as notably expressed by the opposition of Menger's disciple Friedrich Hayek to Keynesianism.

When compared with the methodological scruples displayed in these social science discussions of external validity, the natural sciences seem downright speculative. Take biological research done under the Neo-Darwinian paradigm. Whereas Charles Darwin himself was interested in reconstructing the Earth's unique natural history, Neo-Darwinists are mainly concerned with demonstrating evolution by natural selection as a universal process. Thus, since the 1900s the center of gravity of biological research

has migrated from the field to the laboratory. In Neo-Kantian terms, this results in a tension between idiographic and nomothetic modes of inquiry. However, Neo-Darwinian biologists rarely register the tension as such. Rather, they routinely treat contemporary laboratory experiments as models of causal processes that repeatedly happened to a variety of species in a variety of settings over millions, if not billions, of years. But if Milgram's or Zimbardo's experiments fail to model causal processes across human cultures and human history, there should perhaps be even less grounds for believing that Neo-Darwinian experiments succeed in their even more heroic generalizations across species and eons.

Whereas most threats to validity pertain to the generalization of a single research event, the phrase can be understood in both more macro and micro terms. Macro-threats to validity derive from what logicians call the fallacies of composition and division. In other words, properties of the parts might be invalidly projected as properties of the whole and vice versa. Economists following Menger's lead were especially sensitive to the fallacy of composition, because individuals who act on their own sense of self-interest routinely generate both positive and negative unintended consequences that constrain a society's subsequent range of action. Conversely, it is equally fallacious to infer that a society's overall strength or weakness is caused by some particular strength or weakness in each or most of its members. Thus, those who stress ideological change as a vehicle of social reform are often guilty of the fallacy of division. But at the ultimate microlevel—namely, inferences that a single individual makes from psychological states to epistemological judgments—a still underappreciated source of invalidity is the tendency to use the ease and/or vividness of one's response or recall as a basis for inferring its representativeness of a larger reality it purports to address.

*Steve Fuller*

***See also*** Concurrent Validity; Construct Validity; Content Validity; Criterion Validity; Ecological Validity; External Validity; Internal Validity; Predictive Validity; Research; Threats to Validity; Validity of Measurement; Validity of Research Conclusions

**Further Readings**

Brannigan, A. (2004). *The rise and fall of social psychology: The use and misuse of the experimental method.* New York: Aldine de Gruyter.

Campbell, D.T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Fisher, R. (1925). *Statistical methods for research workers.* Edinburgh, UK: Oliver and Boyd.

Fuller, S. (1988). *Social epistemology.* Bloomington, IN: Indiana University Press.

Fuller, S. (1993). *Philosophy of science and its discontents* (2nd ed.). New York: Guilford.

Fuller, S. (2008). *Dissent over descent.* Cambridge, UK: Icon.

Fuller, S., & Collier, J. (2004). *Philosophy, rhetoric and the end of knowledge* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Habermas, J. (1981). *A theory of communicative action.* Boston: Beacon Press.

Machlup, F. (1978). *Methodology of economics and other social sciences.* New York: Academic Press.

Sober, E. (2008). *Evidence and evolution.* Cambridge, UK: Cambridge University Press.

Zijderveld, A. C. (2006). *Rickert's relevance: The ontological nature and epistemological functions of values.* Leiden, the Netherlands: Brill.

Ziliak, S. T., & McCloskey, D. N. (2008). *The cult of statistical significance: How the standard error costs us jobs, justice, and lives.* Ann Arbor, MI: University of Michigan Press.

## Thurstone Scaling

From the perspective of a person whose attitudes, behaviors, or knowledge are being assessed with a survey of some sort, an attitude scale might seem little more than a series of questions or statements (stimuli) to be answered. For a researcher, however, the instrument used to gather these data is the end product of a process involving considerable reflection on a psychological construct of interest and culminates in a grouping of items that provide material information about people's beliefs and opinions. The process of psychological scaling itself is the measurement and quantification of attitudes, attributes, or traits, and many approaches to this process are available to researchers that vary with respect to both theoretical underpinnings and how they are applied in practice. Among

these is the Thurstone approach, which can be counted among the first formal techniques for attitude measurement.

First theorized by Louis L. Thurstone in the 1920s, Thurstone scaling locates stimuli on a psychological dimension of interest, and then, as individuals respond to the statements that make up a psychological instrument, those people are also located on the continuum of the construct being measured. In this way, Thurstone scaling is typically referenced as an example of the two-step process known as stimulus then person scaling.

At the outset, Thurstone's methodology began with the law of comparative judgment and led to the development of the methods of equal-appearing intervals and successive intervals (the latter was developed in collaboration with Allen Edwards). In this entry, a general overview of Thurstone's methods is provided, and modern applications of the fundamentals of the Thurstone scaling approach are described as well.

## Two-Step Scaling and the Thurstone Approach

As in any scaling exercise, the beginning step is to operationalize the construct of interest and several statements that help to define that attitude such as beliefs, feelings, and/or behaviors are written. These must be reviewed by judges (in this case, the participants in a scale development study) in the first part of the two-step approach to scaling. The goal of this step is not to obtain the actual opinions of the participants toward the construct of interest. Rather, what are needed here are their judgments about the favorableness of those statements on that construct, to identify how items differentially reflect specific levels of an attitude or trait. Typically, many more statements than are strictly necessary are written, as some winnowing of statements might occur in the scale development process to create the final measurement instrument from the assortment of scaled stimuli.

Next, taking the statement ratings across judges, the stimuli can then be sequenced on a subjective continuum from least to most favorable. Positions on this continuum should span the full range from strongly negative through neutral to strongly positive. Then, for the last piece of scale assembly, the statements to be included are deliberately sequenced in no particular order; that is to say, they are not sorted by their scale values.

Once the scale is assembled in this manner, the second step of scaling occurs through administration of the scale itself, as people's endorsements (or not) of the statements on the scale allow for positioning the respondents to the survey on the attitude continuum. The attitude score for a person is the mean (or median) scale score of the stimuli that the individual agreed with. An important point to be made here is that the actual *scaling* portion of Thurstone scaling in fact references three distinct approaches to the judges' task of rating the stimuli, as described below.

### *The Law of Comparative Judgment*

This first approach to scaling the statements by way of judges' ratings is the law of comparative judgment, which more aptly describes a model rooted in the notion of comparisons between two stimuli at a time. This was originally conceptualized as involving physical entities (such as handwriting specimens), but was subsequently generalized by Thurstone to psychological qualities.

In this method, each entity or stimulus is compared with each and every other item; hence, it is also referenced as pairwise comparisons. As a part of scale development, the comparative judgment technique depends on judges' determinations of which of each pair of stimulus statements belong above the other (e.g., more favorable toward the attribute being measured). The kinds of judgments made are described by Thurstone as any qualitative or quantitative attribute for which we can think "more" or "less" for each specimen.

Once the judges' ratings are made, the scaling task uses proportions to rank statements. Proportionally, the more often a statement is deemed to be more favorable than another determines the distance between the statements; thus, the scale emerges through the rating data based on the perceived differences among the entire collection of items, and it is obtained through the complete series of paired comparisons. Statistically, the scale value for each statement is obtained using these proportions and the standard errors of observations. An illustrative example of this computation is provided in Thurstone's writings on paired

comparisons, particularly his 1927 article in which 266 students were asked to judge 19 crimes on a scale of seriousness. Rape, then homicide, was judged to be the most serious, whereas vagrancy was considered the least-serious offense.

Recognizing that a scale is an artificial construct created through the presentation of stimuli that are not fixed but might exhibit small fluctuations in their values from presentation to presentation, one assumption in Thurstone scaling is normality, in the sense that the distribution of judgments on the judgment dimension is normal for each stimulus and centered on the stimulus's scale value, not that the stimuli themselves are normally distributed. Indeed, Thurstone's *discriminal process* can be viewed as a conceptual analog of a random variable in current research methods.

It should be noted that comparative judgment in some applications can be cumbersome in practice depending on the pool of stimulus items that must be reviewed in this pairwise approach, because with $n$ statements, the number of pairs is computed as $[n(n - 1)]/2$. If a final scale of 20 to 25 items is needed, then the number of comparisons is considerable. A 25-item scale minimally requires 600 paired judgments from each judge, if exactly 25 stimuli are tried out; of course, sometimes scale developers expect a scale to undergo some refinement by way of item selection and will try out more stimuli than they might use in practice. Thus, if the comparative judgments technique is used, it translates into more statements and more judgments. Across all judges, the proportion that a given statement is judged more favorable than its mate in each comparison allows the scale developer to position items on the scale.

### *Method of Equal-Appearing Intervals*

The scaling task for this Thurstone scaling method is considerably different from that in the comparative judgment approach. This method is a sorting task in which the judges are instructed to consider each stimulus independently and then assign a value to each stimulus on a scale of 1 to 11, with one end of the scale set to "less favorable" and the other end anchoring "more favorable." Again, the judgment of interest is not the judges' own agreement/disagreement with the stimuli, but rather it is the extent to which each stimulus indicates that the future respondent possesses a large or small quantity of the quality being measured. For example, a rating of 1 for an item to be used on an instrument measuring misogyny might indicate that the person has a low amount of negative feelings toward women, whereas an 11 would correspond to a high level of misogynistic attitude.

By arranging the stimuli in this way, each judge creates 11 "piles" or intervals, and it typically results in intervals that to each judge appear to be more or less equal in width. There is some question whether judges are instructed that the intervals are equal or not, as Thurstone's first writings on this method are silent on the issue but in later works Thurstone is clear that it was an explicit instruction. Whether overtly said or not, the assumption underlying this approach is that the intervals are equal, which permits the scale developer to set the width of each interval to 1 and to assign a sequential value to each interval, from 1 to 11. Obtaining the median score of the judges' ratings for each stimulus then completes the scaling task.

### *Method of Successive Intervals*

Successive intervals scaling was subsequently developed as a response to the imperfect assumption underlying the equal-appearing intervals approach; as it turned out in practice, the scale intervals were not equal, with more stimuli located at the ends of the continuum and fewer in the middle. This scaling method follows much of the same path as equal-appearing intervals, in that the rating/sorting task is the same. However, instead of adhering to an advance belief that the intervals are equal to one another these intervals can be empirically estimated by using Thurstone's earlier assumption about the normality of the judgments for each stimulus. This method was first described by Milton Saffir, and later by Allen Edwards and Thurstone, among others.

As with equal-appearing intervals, the successive intervals approach begins with the simple proportions of interval judgments for each item. Next, the cumulative proportions of judgments for all items are provided. These cumulative proportions can then be represented by $z$ score values via the use of a lookup table to create a matrix of $z$

scores. The $z$ scores in these cells are in effect boundaries of the intervals, and by subtracting the values in adjacent cells, it is possible to estimate interval widths. Whereas the equal-appearing intervals method sets this value as 1, in reality, these widths can vary considerably.

When these data are taken with the data for all items in the stimulus pool for a given instrument, by working in columns, the mean interval width can be computed for each interval, thus defining a scale for the intervals. To obtain the scale values for individual items, the scale must first be centered. A simple way to do this is to compute the mean of interval widths for each row and then use the row means (one for each stimulus) to work down and obtain a mean of the row means. That value can then be set as the zero point for the scale; then, by subtracting the row mean from the mean of row means, the location for each item on the scale can also be determined. Other extrapolation methods for computing interval widths and stimulus locations were detailed by Edwards in the 1950s.

## Item Selection

In discussing Thurstone's scaling methods, an important consideration in scale development that warrants mention is item selection. Given the nature of the sorting task in both the equal-appearing and successive intervals approaches and that it involves judges' ratings for different stimuli, the stimuli that are preferred for inclusion on the final version of the instrument are those with a high level of agreement among judges. Greater dispersion among the judgments might be an indicator that the stimulus is ambiguous to the judges (and potentially the respondents as well). Thurstone and Ernest Chave in collaboration suggested that the interquartile range could be used as an indicator of the level of dispersion.

A second consideration in this regard is to ensure that the items included on a scale appropriately differentiate between respondents with low and high amounts of the quality of interest. If a stimulus does not discriminate between respondents in this way, then it does not contribute unique measurement information about respondents to help locate them on the continuum of the attribute being assessed. For scale developers, item discrimination in the Thurstone sense can be evaluated by looking at the probability of agreement with the stimulus items conditional on attitude scores across the scale. This operative characteristic curve can be plotted graphically and reviewed.

## Modern Applications

The Thurstone scaling approaches detailed here have been in use since the 1920s, and in that time, these methods have helped shape the practice of modern psychometrics significantly. Beginning with his 1925 paper "A Method of Scaling Psychological and Educational Tests," Thurstone's thinking about attitude measurement was also applied to the realm of educational assessment, considering test items that can be graded as right or wrong, and for which separate norms are to be constructed for successive age or grade groups.

Thurstone's *absolute method* for scaling items permitted scale developers to create scales that were independent of data, specifically in that the scale is independent of the unit selected for the raw scores and the shape of the distribution of the raw scores. His 1925 paper illustrated how the method could be used for scaling tests across successive age groups (a practice now commonly referred to as vertical equating). This approach assumes that a normal distribution of the construct being measured within an age/grade group, and each item can be placed on the scale such that the probability of correct responses is equal to the probability of incorrect responses (in effect, as described by Wendy Yen, the $p$ value for each item is transformed to a $z$ value using the inverse cumulative normal function). Thurstone's methods were employed a great deal by many educational testing agencies in the second half of the 20th century.

These theoretical developments that occurred throughout the 1920s to conceptualize persons and items on the same scale with regard to an underlying psychological construct (often referred to as a latent trait in the item response theory [IRT] context) helped to pave the way for modern psychometrics, and this is particularly apparent in the connections that can be found between Thurstone's work and IRT. The relationship between the Thurstone approach to vertical scaling and scaling using a three-parameter IRT model is strong, particularly with respect to the normality

and equal interval assumptions found in Thurstone scaling. Thurstone's philosophy of measurement is also considered to have contributed to the foundation of the Rasch IRT tradition for assessment.

Empirical analyses comparing the distribution of performance across grade levels as produced by Thurstone and IRT scaling have also been carried out. In this research, one quality of Thurstone's scaling that was specifically studied was the finding that variability of performance tended to increase with age under Thurstone's method, in contrast to IRT, where decreasingly variability was often observed. The results indicated that the two approaches yielded reasonable consistency in growth trends. Other researchers who examined Thurstone scaling in relation to one and three-parameter IRT models also found increased variability as grade levels increased.

## Conclusion

Thurstone scaling represents a family of approaches to quantifying psychological qualities, and the idea that "attitudes can be measured," as Thurstone wrote in 1928, was a significant part of a movement that transformed not only psychology but also educational assessment. The principles underlying his scaling theory emerged at a time when the idea of objective mental measurement itself was being formalized by Thurstone and others, and thus his contributions in this regard to modern psychometrics are undeniable.

In the intervening years since these methods were first established, Thurstone himself and others have identified some modifications and limitations to them. For example, as noted previously, the scope of the task in the paired comparisons method is one obstacle to its use when the number of stimuli might be large. The main underlying assumption of equal-appearing intervals has been found to not be supported through research, and therefore the results obtained through that method might not be as robust as they might have appeared. Last, whereas Thurstone approaches account for random variation through the idea of the discriminal process, some questions have been raised about the potential for systematic variation to emerge through the judges' ratings. Those ratings themselves might be systematically biased by the judges' own attitudes, and the possibility exists that those biases could impact the scaling in a way that might not be accounted for by Thurstonian logic.

Ultimately, as an approach to scaling, Thurstone's work serves as one strategy among many that offers scale developers with a process for formally quantifying information that was once not thought possible. The balance of sophistication and elegance in his original conceptualizations is considerable, and Thurstone scaling methods are valued for their substantial contribution to psychometrics.

*April L. Zenisky*

***See also*** Distribution; Item Response Theory; *p* Value; Raw Scores; Theory of Attitude Measurement; *z* Score

### Further Readings

Blischke, W. R., Bush, J. W., & Kaplan, R. M. (1975). Successive intervals analysis of preference measures in a health status index. *Health Services Research, 10,* 181–198.

Edwards, A. L. (1952). The scaling of stimuli by the method of successive intervals. *Journal of Applied Psychology, 36,* 118–122.

Edwards, A. L., & Thurstone, L. L. (1952). An interval consistency check for scale values determined by the method of successive intervals. *Psychometrika, 17,* 169–180.

Saffir, M. A. (1937). A comparative study of scales constructed by three psychophysical methods. *Psychometrika, 2,* 179–198.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *The Journal of Educational Psychology, 16,* 433–450.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34,* 273–286.

Thurstone, L. L. (1927). Psychophysical analysis. *American Journal of Psychology, 38,* 368–389.

Thurstone, L. L. (1927). The method of paired comparisons for social values. *Journal of Abnormal and Social Psychology, 21,* 384–400.

Thurstone, L. L. (1927). The unit of measurement in educational scales. *Journal of Educational Psychology, 18,* 505–524.

Thurstone, L. L. (1928). Attitudes can be measured. *The American Journal of Sociology, 33,* 529–554.

Thurstone, L. L., & Chave, E. J. (1929). *The measurement of attitudes.* Chicago: University of Chicago Press.

Yen, W. M. (1986). The choice of scale for attitude measurement: An IRT perspective. *Journal of Educational Measurement, 23,* 299–325.

# Time-Lag Study

A time-lag study examines the responses of different participants of similar age at different points in time. Time-lag is one of the three methods used to study developmental and generational change. The other two methods are a cross-sectional study (which examines participants of different ages at one point in time) and a longitudinal study (which examines the same participants as they age). This entry first examines the types of differences these methods assess; then, it describes the possible confounds and the procedures to follow to perform a time-lag study. Last, this entry briefly discusses the future of time-lag studies.

## Differences

These methods assess three types of differences: age differences (a result of development), generational differences (a result of generational succession), and time period (a result of historical events that affect all generations equally). Longitudinal studies, which follow a group of people of the same age over time, can determine age and time period differences (but not generation, as the participants are all the same generation). Cross-sectional studies collect data at only one time; thus, any differences could be a result of age or generation (but not time period, as the time period is the same). Time-lag studies examine people of the same age at different points in time; thus, differences could be a result of the generation or time period (but not age, as age is the same). Sequential designs are the best way to separate age, time, and cohort effects. For example, a time-lag study might study several age groups over time. However, a change across all ages, usually interpreted as a time effect, could still be a cohort effect if the cohort change has been continuing for many years. For this reason, some authors have concluded that it is impossible to completely separate age, cohort, and time effects even with sequential designs. Most researchers agree that it is most important to separate age and cohort effects, as time period is a less important effect.

Thus, time-lag studies are the best method for examining generational (or birth cohort) differences. Most research and theorizing suggest that historical events and cultural trends have the most impact on the attitudes of children and adolescents; thus, generational effects are stronger than time period effects. For example, the large changes in sexual attitudes and behavior found in a time-lag study headed by Brooke Wells are primarily driven by generational change, as people tend not to change their views on sexuality much once they are past young adulthood.

Time-lag studies illustrate how historical changes in culture affect individuals. Recent research and theory in psychology have recognized that environments vary between countries and regions, producing differences in personality, emotion, perception, and behavior. Environments vary over time and generations in a similar way (for example, the United States in the 1950s was a different environment than the United States in the 2000s). The time-lag method makes it possible to study generational change scientifically, separating the effects of age and generation (as a cross-sectional study, with data collected at the same time, cannot do).

Some time-lag studies examine only two or three data points. Other studies, such as Monitoring the Future, The American Freshman, and the General Social Survey, collect data from participants every year for many years. For example, Monitoring the Future data show that 24% of high-school seniors in 1976 anticipated receiving a graduate degree, which jumped to 50% by 2000. Time-lag methods can be used for health measures as well; one time-lag study found that tooth loss among older people was primarily explained by cohort rather than age or time.

A method called cross-temporal meta-analysis modifies the time-lag method by gathering mean responses on psychological scales from journal articles and unpublished dissertations using many data points (always more than 10, and usually 50 or more). These studies examine data separately by age group (e.g., college students and middle school students), analyzing whether the means differ by the year of data collection. For example, one study examined whether scores on the Narcissistic Personality Inventory had changed over time. The analysis showed that college students' NPI scores were significantly

higher in 2006 than they were in 1979. This was a time-lag study because the participants were the same age (college students) but completed the scale at different points in time (sometime between 1979 and 2006).

## Confounds

Like any research study, a time-lag study must consider possible confounds. If the time of data collection and location are correlated, location (and the accompanying differences in the population) rather than time might cause any differences or suppress differences. Confounds by location are more problematic if there are only a few data points, introducing the possibility of location and year being perfectly confounded—that is, all the earlier data point(s) were collected at one location, and the later data point(s) at another location. For example, Jerry Burger's partial replication of Milgram's obedience experiment collected data on an ethnically diverse sample of men and women from California in 2006, which was different from Stanley Milgram's nearly all White sample of men only from New Haven, Connecticut, in the early 1960s. Thus, the differences, or in this case lack thereof, might have been caused by sample differences in gender, ethnicity, or location. Another study found that college students' narcissism scores were flat in 8 samples from University of California campuses; however, the first 2 samples (in 1982 and 1996) were from one campus, and the later samples (2002–2007) were from another. Thus the change in location might have obscured any change. When the 2002–2007 samples from one campus were examined, they showed an increase in narcissism scores similar to that found in a nationwide analysis.

Changes in the population sampled can also cause problems. More women have attended college over time, so an increase in the mean score of anxiety measures for college students could have been caused by more women in the samples, as women score higher on measures of anxiety. In a time-lag analysis of anxiety scores, men's and women's data were analyzed separately and still showed the effect, so in this case, there was a cohort effect apart from the confound.

## Procedures

There are three primary ways to do a time-lag study. The first requires patience: One can collect data from a sample (preferably of the same age or with many people from every age group) and then wait to collect more data—either every year or after the desired interval. Researchers will sometimes compare recent data with data they have collected earlier in their careers. This strategy has the advantage of openness of research question: One can ask the questions desired. Second, researchers can draw from one of the large time-lag studies such as the General Social Survey. The advantage is that the data have already been collected back in time; the disadvantage is one must make do with the questions in the survey. Another advantage is that many of these large projects attempted to obtain nationally representative samples.

The third choice is to conduct a cross-temporal meta-analysis, gathering mean scores on psychological scales. This type of study can be conducted at one time, but gives a historical view through other researchers' reports. There is some restriction on the characteristics to be examined; the scale(s) must have been used by many researchers, and the mean scores must have been reported by many of them, over a good span of time (at least 10 years, and preferably 20 or more). To do a cross-temporal meta-analysis, the first step is to choose a scale measuring the desired concept and search the Web of Science citation index or a similar database to determine whether enough studies have been published that cite the main source for the scale (at least 100 is a good rule of thumb). One then obtains the full text of these articles (and of dissertations using Digital Dissertations) and records the mean score, standard deviation, year of data collection, $n$, and characteristics of the sample (gender, ethnicity, location, etc.). Analyses are performed split by age group (e.g., middle school, high school, and college students), as a time-lag study must use respondents of the same age. A regression is then performed with year as the independent variable, mean as the dependent variable, and $n$ of each sample as the WLS weight (WLS stands for Weighted Least Squares, but most statistical programs list it simply as WLS). If the regression is significant, then an additional step must be taken to calculate the effect size: Multiply the

unstandardized B by the number of years examined (e.g., 20 for 1985–2005) and divide by the mean standard deviation from the samples. This technique avoids the ecological fallacy (also known as alerting correlations), which occurs when an effect size relies on the standard deviation of a group of means (which is smaller) rather than the standard deviation of a group of individuals (which is larger). Because this method uses the standard deviation given in the samples—and thus the standard deviation of a group of individuals—it avoids the ecological fallacy.

Time-lag studies can also examine how traits and attitudes differ along with specific aspects of an historical period. For example, to determine whether children's anxiety is correlated with the overall divorce rate, social statistics (e.g., the divorce rate) are used in place of the year of data collection in an analysis. These analyses cannot prove causation, but they can show what social trends occur at the same time as shifts in personality traits and attitudes. Analyses can also use the responses of other time-lag studies; for example, materialism measured by the American Freshman survey is correlated across time with symptoms of depression and anxiety as measured by the Minnesota Multiphasic Personality Inventory.

### Future Directions

Although more research is emerging on generational change, a great deal remains to be done. Exactly how cultural change occurs is a fascinating and mostly unanswered question. The studies that attempt to address cultural and generational change in the future will most likely rely on the time-lag method.

*Jean M. Twenge*

***See also*** Cross-Sectional Design; Effect Size, Measures of; Longitudinal Design; Regression to the Mean; Standard Deviation

#### Further Readings

Burger, J. M. (in press). Replicating Milgram: Would people still obey today? *American Psychologist.*

Reynolds, J., Stewart, M., MacDonald, R., & Sischo, L. (2006). Have adolescents become too ambitious? High school seniors' educational and occupational plans, 1976 to 2000. *Social Problems, 53,* 186–206.

Schaie, K. W. (1965). A general model for the study of developmental problems. *Psychological Bulletin, 64,* 92–107.

Twenge, J. M. (2006). *Generation me: Why today's young Americans are more confident, assertive, entitled—and more miserable than ever before.* New York: Free Press.

Twenge, J. M., Konrath, S., Foster, J. D., Campbell, W. K., & Bushman, B. J. (2008). Egos inflating over time: A cross-temporal meta-analysis of the Narcissistic Personality Inventory. *Journal of Personality, 76,* 875–901.

Wells, B. E., & Twenge, J. M. (2005). Changes in young people's sexual behavior and attitudes, 1943–1999: A cross-temporal meta-analysis. *Review of General Psychology, 9,* 249–261.

## Time-Series Study

Time-series analysis (TSA) is a statistical methodology appropriate for longitudinal research designs that involve single subjects or research units that are measured repeatedly at regular intervals over time. TSA can be viewed as the exemplar of all longitudinal designs. TSA can provide an understanding of the underlying naturalistic process and the pattern of change over time, or it can evaluate the effects of either a planned or unplanned intervention. The advances in information systems technology are making time-series designs an increasingly feasible method for studying important psychological phenomena.

Modern TSA and related research methods represent a sophisticated leap forward in the ability to analyze longitudinal data. Early time-series designs, especially within psychology, relied heavily on graphical analysis to describe and interpret the results. Although graphical methods are useful and provide important ancillary information, the ability to bring a sophisticated statistical methodology to bear has revolutionized the area of single-subject research.

TSA was developed more extensively in areas such as engineering and economics before it came into widespread use within social science research. The prevalent methodology that has developed and been adapted in psychology is the class of

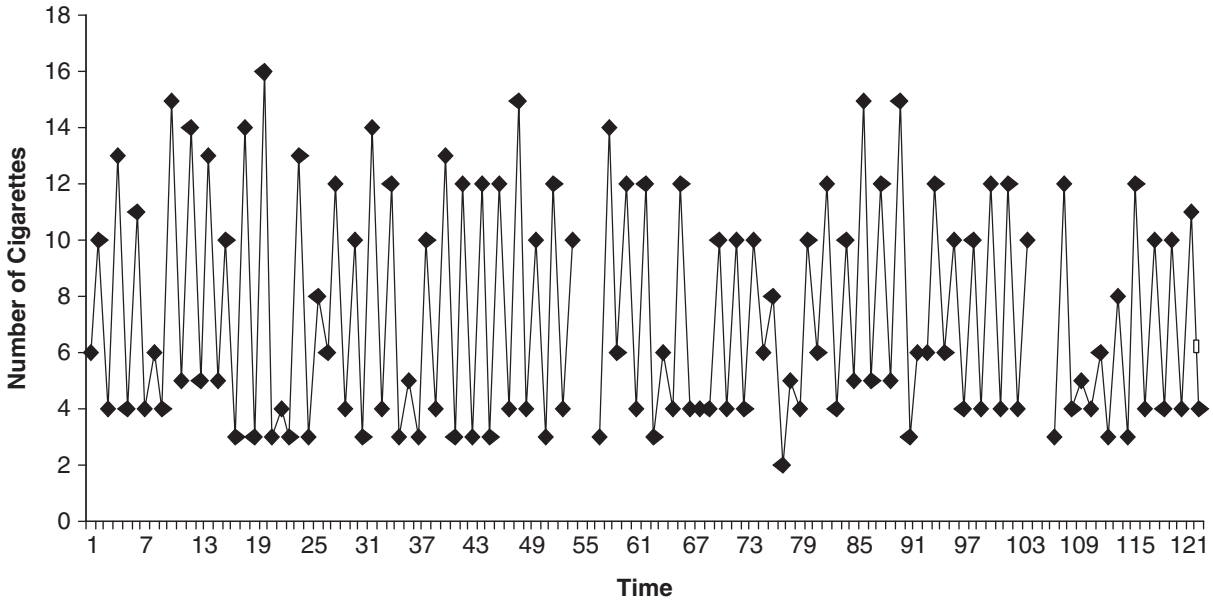



**Figure 1** Smoking Behavior Measured on 124 Occasions: An Example of Time-Series Data for a Single Individual

models known as Autoregressive Integrated Moving Average (ARIMA) models. TSA requires the use of high-speed computers; the estimation of the basic parameters cannot be performed by precomputer methods.

A major characteristic of time-series data is the dependency that results from repeated measurements over time on a single subject or unit. All longitudinal designs must take dependency into account. Dependency precludes the use of traditional statistical tests because they assume the independence of the error. ARIMA models have proven especially useful because they provide a basic methodology to model the dependency from the data series and allow valid statistical tests. This entry discusses several aspects of TSA, including its application in research, modeling procedures, weakness and confounders, and the use of telemetrics in TSA.

## Research Applications

As the methodology for TSA has evolved, there has also been increasing interest among applied researchers. Many behavioral interventions occur in applied settings such as businesses, schools, clinics, and hospitals. More traditional between-subject research designs might not always be the most appropriate, or in some instances, these designs can be very difficult to implement in such settings. In some cases, data appropriate for TSA are generated on a regular basis in the applied setting, like the number of hospital admissions. In other cases, a complete understanding of the process that can explain the acquisition or cessation of an important behavior might require the intensive study of an individual during an extended period of time. Advances in information systems technology have facilitated the repeated assessment of individuals in natural settings. In addition, there has been increasing concern about the appropriateness of group methods based on the ergodic theorems. The ergodic theorems state that an analysis of the interindividual variation yields the same results as an analysis of intraindividual variation only if the trajectory of each subject obeys the same dynamic laws (i.e., the same autocorrelation structure), and each individual trajectory has the same statistical characteristics (i.e., equal mean level and temporal pattern).

Several important classes of research questions can be investigated using time-series data. The first class involves using ARIMA modeling to investigate the naturalistic process of change across time. Investigations of this type focus on the dependency parameters and attempt to identify the underlying nature of the series. Figure 1 illustrates time-series data for an individual smoker with no intervention attempted. The data are the number of cigarettes smoked during an 8-hour period for a single individual (two periods per day).

A second important class of research questions involves the analysis of the effects of an intervention that is applied to an individual subject or unit. Such an investigation is commonly referred to as an *interrupted time-series analysis*. The *interruption* refers to the intervention that is applied at some fixed point in a process. Repeated measurements are taken before and after the intervention to evaluate the effects of the intervention. Such investigations can be very useful in trying to understand causality within the process and as a result of the intervention.

A third class of research questions involves the pattern of change over time. These patterns can be varied, and some questions that can be investigated in this context include the following: (a) Are the effects of intervention temporary or permanent, (b) does the intervention cause a change in the "slope" of the behavior process as well as the overall level, (c) does the intervention cause a change in any cycle that is present in the underlying behavior process, (d) does the intervention cause the variance to change, and (e) does the intervention cause a change in the nature of the dependency?

A fourth class of research questions involves forecasting future values of the series. This strategy is primarily used in economics.

## Caveats

Several difficulties and weaknesses associated with standard TSA must be recognized. First, generalizability should not be inferred from a single study. The researcher needs to engage in systematic replication to demonstrate generalizability. Second, the traditional measures employed in cross-sectional studies in many content areas might not be appropriate for time-series designs. For TSA, the best measures are those that can be repeated a large number of times on a single subject at intervals of short duration. Third, many equally spaced observations are required for accurate model identification, which is a necessary step for identifying basic processes.

## ARIMA Modeling Procedures

TSA, within the ARIMA model framework, involves two important steps that can vary in importance depending on the goals of the study. The first step is *model identification,* in which the researcher tries to identify which underlying mathematical model is appropriate for the data. Model identification focuses on dependency parameters.

The ARIMA model represents a family of models characterized by three parameters ($p$, $d$, and $q$) that describe the basic properties of a specific time-series model. The value of the first parameter $p$ denotes the order of the autoregressive component of the model. If an observation can be influenced only by the immediately preceding observation, then the model is of order one. If an observation can be influenced by the two immediately preceding observations, then the model is of order two, and so on. The value of the second parameter $d$ refers to the order of differencing that is necessary to stabilize a nonstationary time series. A process is described as nonstationary when the values do not vary about a fixed mean level; rather, the series might first fluctuate about one level for some observations and fluctuate about a different level. The value of the third parameter $q$ denotes the order of the moving average component of the model. Again, the order describes how many preceding observations must be taken into account. Higher order models ($p$ or $q > 3$) are rare in the behavioral sciences.

Dependence is assessed by calculating the values of the autocorrelations among the data points in the series. In contrast to a correlation coefficient, which is used to estimate the relationship between two variables measured at the same time on multiple subjects, an autocorrelation estimates the relationships within one variable measured at regular intervals over time on only one subject.

The direction of dependency in a time series refers to whether an autocorrelation is positive or negative. The direction can be determined with

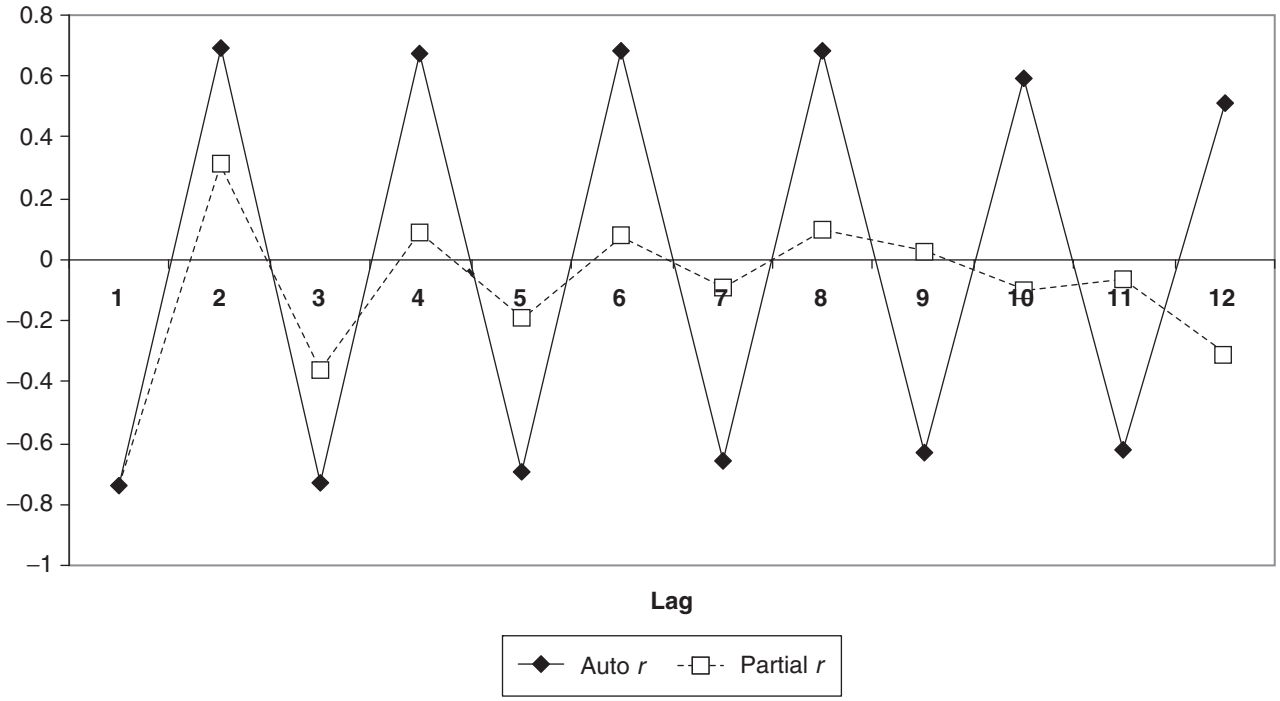



**Figure 2** Correlogram of the Autocorrelations and Partial Autocorrelations for the Smoking Data Example

a high degree of accuracy when there is strong dependency in the data and the direction has clear implications. When the sign of the autocorrelation is negative, a high level for the series on one occasion predicts a lower level for the series on the next occasion. When the sign is positive, a high level of the series on one occasion predicts a higher level on the next occasion.

The calculation and interpretation of the pattern of the related partial-autocorrelations calculated at each lag is employed as a second diagnostic step to aid in the identification of the specific ARIMA model. Partial-autocorrelations are mathematically complex and will not be formally defined here. Figure 2 illustrates the autocorrelations and partial-autocorrelations for the data from Figure 1.

Figure 3 illustrates four different types of models using computer-generated data for 60 observations before an intervention and 60 observations after the intervention. The first graph (a) represents an ideal interrupted time-series example initially at level = 5.0 with no error and an immediate change in level of 2.0 units at the time of intervention. The next three graphs represent an order 1 autoregressive model [ARIMA (1, 0, 0)]. The second graph (b) adds random error with a variance of 1.00. There is no autocorrelation in this model. The third graph (c) is the same change in level and error variance as (b) but with a large *negative* autocorrelation ($-0.80$). The fourth graph (d) is a model with the same change in level and error variance but with a large *positive* autocorrelation ($+0.80$). The impact of dependency can be easily observed. The negative dependency results in an exaggerated "saw tooth" graph with increased apparent variability. The positive dependency results in a smoother graph with apparent decreased variability. The inclusion of an intervention effect (the change in level) illustrates how difficult it is to determine whether an intervention had an effect by visual inspection alone.

## Interrupted Time-Series Analysis

Often, the goal of research with single subjects or units is to determine the efficacy of a specific

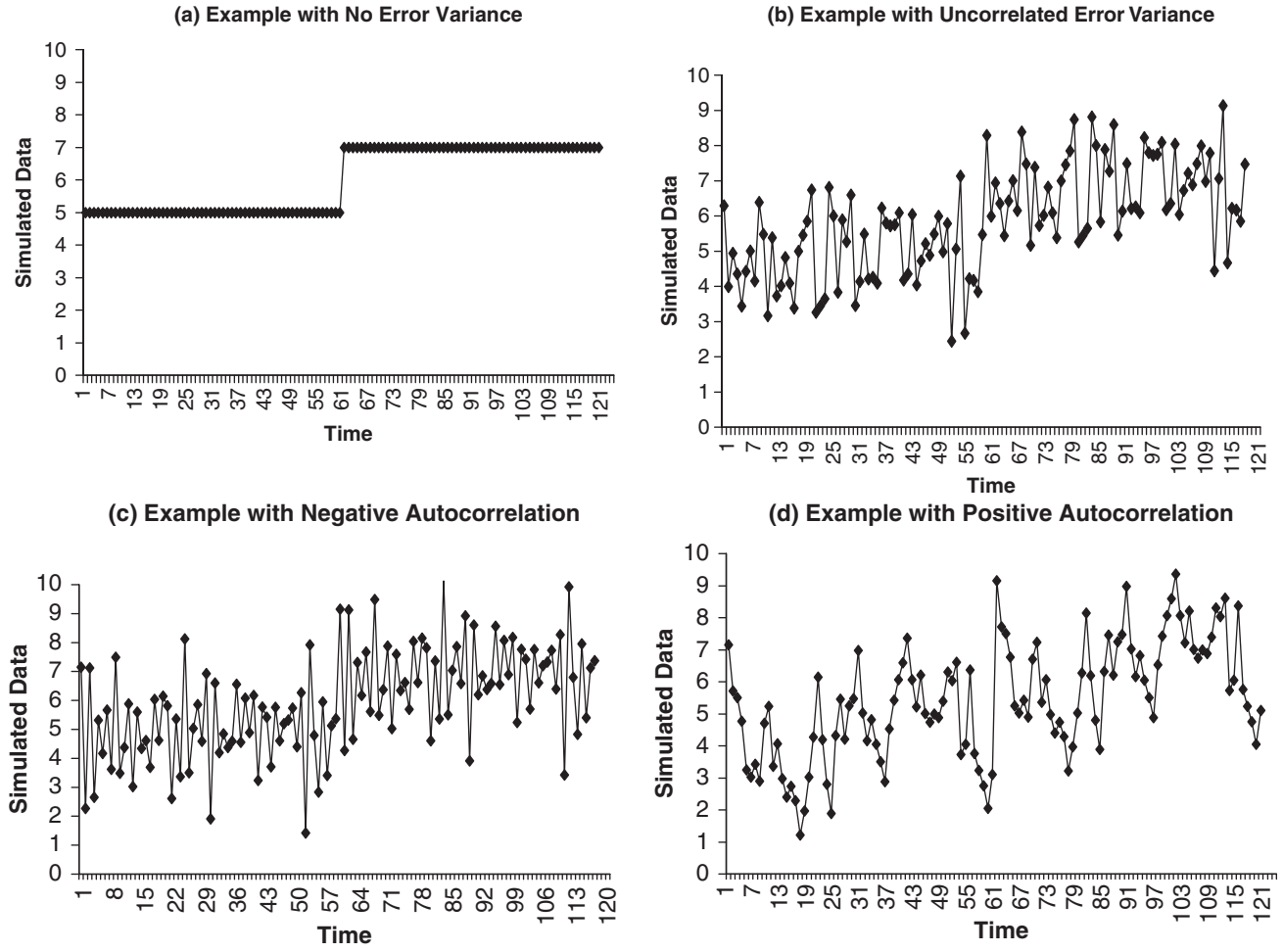



**Figure 3** Computer-Generated Data for ARIMA (1, 0, 0) Models for Level = 5.0 and Change in Level = 2.0 for Four Time Series Illustrating Different Degrees of Dependency

intervention. This can be accomplished by employing various techniques that fall under the nomenclature of interrupted time-series analysis. A simple example of an interrupted TSA is a design that involves repeated and equally spaced observations on a single subject or unit followed by an intervention. The intervention would then be followed by additional repeated and equally spaced observations of the subject or unit. The intervention could be an experimental manipulation, or it could be a naturally occurring event. To determine whether the intervention had an effect, preprocessing of the data series to remove the effects of dependence is needed. The actual statistical analysis used in an interrupted TSA employs a general linear model analysis using a generalized least-squares estimator.

If the intervention effect is found to be statistically significant, then a related question concerns an evaluation of the nature of the effect. One great advantage of TSA is the ability to assess the pattern of the change over time, which can involve both a change in the mean level of a measured dependent variable and/or a change in the slope over time of the dependent variable.

The most common methodology for interrupted TSA is the Box-Jenkins procedure, which is a two-step process. The first step is determining which ARIMA model fits the data, and the second step is the analysis of the effects of the interruption.

Model identification is necessary to determine the specific transformation matrix to be used to remove the dependency from the data series.

The general linear model (GLM) is an approach to data analysis that includes many familiar statistical procedures as special cases. In algebraic terms this can be expressed as

$$\mathbf{Z} = \mathbf{Xb} + \mathbf{a}. \tag{1}$$

In a multiple regression analysis, the $\mathbf{X}$ matrix contains the numeric observations for each of the $p$ predictor variables for the $N$ subjects, the $\mathbf{Z}$ vector contains the criterion scores for the $N$ subjects, the $\mathbf{b}$ vector contains the regression weights, and the $\mathbf{a}$ vector contains the error of prediction and represents the difference between the actual score on the criterion and the predicted scores on the criterion. In an analysis of variance, the $\mathbf{X}$ matrix would consist of indicator variables, such as the numeric values "1" or "0," which indicate group membership, and the $\mathbf{Z}$ vector contains the dependent variable observations.

For both of these examples, the estimates of the parameters are

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}. \tag{2}$$

For TSA, the parameter estimates include a lower diagonal transformation matrix. The exact nature of the transformation matrix is determined by the model identification step. The parameter estimates are now

$$\mathbf{b} = (\mathbf{X}'\mathbf{T}'\mathbf{T}\mathbf{X})^{-1}\mathbf{X}'\mathbf{T}'\mathbf{T}\mathbf{Z}. \tag{3}$$

For example, an ARIMA (1, 0, 0) model with five observations would have the following transformation matrix:

$$\mathbf{T} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ \phi_1 & 1 & 0 & 0 & 0 \\ 0 & \phi_1 & 1 & 0 & 0 \\ 0 & 0 & \phi_1 & 1 & 0 \\ 0 & 0 & 0 & \phi_1 & 1 \end{bmatrix}, \tag{4}$$

which indicates that only the previous observation is necessary to explain the dependency in the data. The GLM can be viewed as a special case of the more general time-series model.

For an interrupted TSA, there are typically four parameters of interest, which include the level of the series (L), the slope of the series (S), the change in level (DL), and the change in slope (DS). The slope parameters represent one of the other unique characteristics of a longitudinal design, the pattern of change over time.

Figure 4 illustrates eight different outcomes for a simple one-intervention design. By inspecting the different patterns of change over time, it can be observed that selecting different points in time for the single assessment would result in different conclusions for four of the examples (c, f, g, and h). For instance, ignoring the slope in (c) would lead the researcher to conclude incorrectly that the intervention was effective. The evolutionary effect (h) is a good example of where the intervention results in a temporary negative effect, perhaps while a response pattern is unlearned, followed by a positive effect. An early assessment would conclude that the treatment had a negative effect; a somewhat later assessment would find no treatment effect, whereas an even later assessment would find a positive treatment effect.

Although it is the most prevalent time-series methodology, the Box–Jenkins approach to intervention analysis suffers from several difficulties. First, gathering the number of data points required for accurate model identification is often prohibitive for research in applied settings. Second, even with the required number of data points, correct model identification is problematic. Third, the method is complex, making applications by the mathematically unsophisticated researcher difficult. Three alternative approaches are described in the next section, all of which attempt to avoid the problematic model identification step.

Dean Keith Simonton proposed a procedure that uses an estimate of the variance-covariance matrix based on pooling the observations across all subjects observed. This approach requires a basic assumption, namely that all series are assumed to be an ARIMA (1, 0, 0) model. While this assumption seems to be theoretically indefensible, empirical investigations indicate that this procedure works well in a wide variety of cases. James Algina and Hariharan Swaminathan have proposed an alternative in which the sample variance-covariance matrix is employed as an estimator for $\mathbf{T'T}$ in the modified least squares

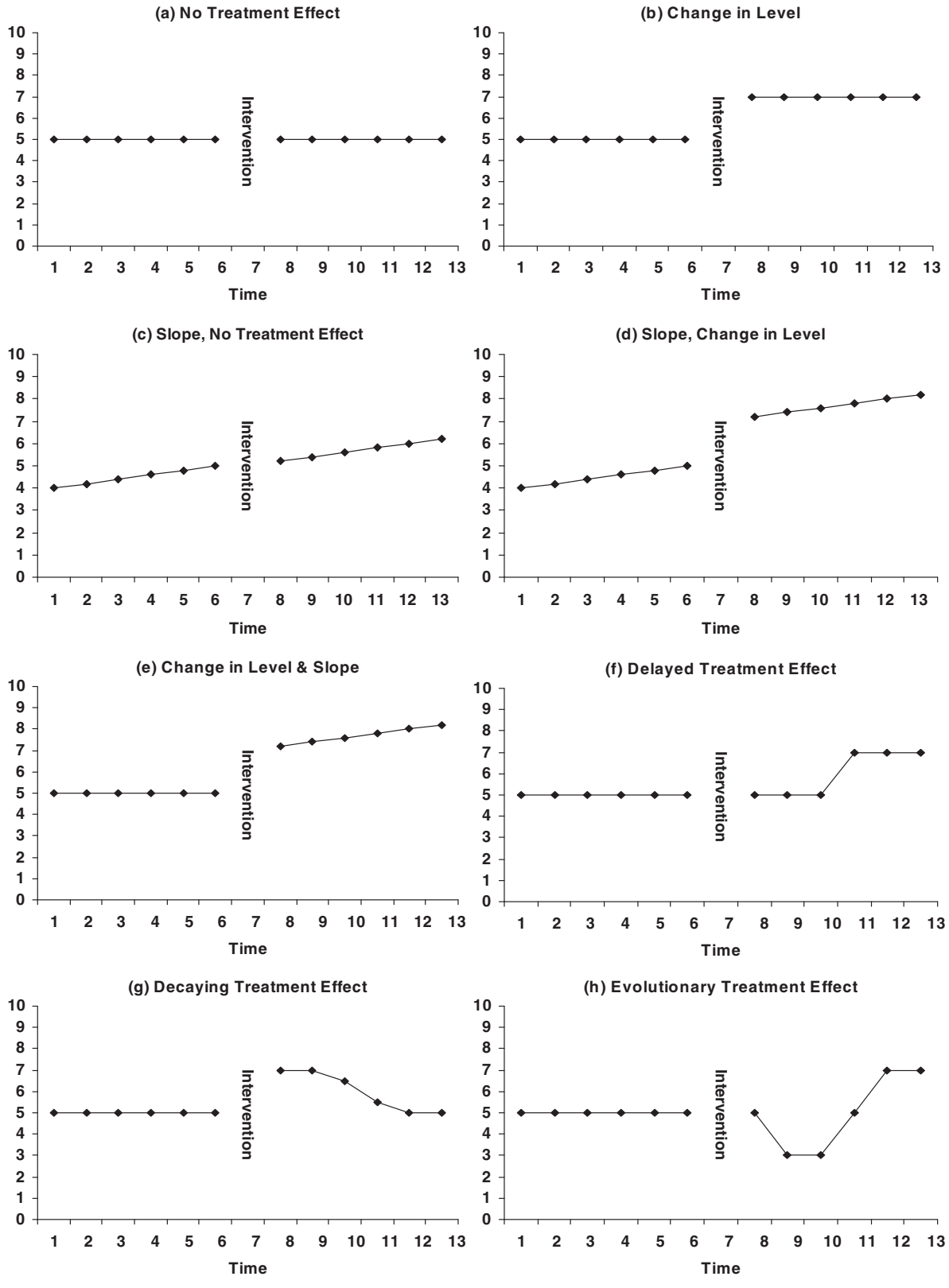



**Figure 4** Examples of Eight Different Patterns of Intervention Effects

solution. This approach, however, requires the assumption that the number of subjects is greater than the number of observations per subject, a condition unlikely to be met in most applied research settings. Wayne Velicer and Roderick McDonald have proposed the use of a *general* transformation matrix with the numerical values of the elements of **T** being estimated for each problem. The rationale is that all transformation matrices **T** have an identical form and use a lower triangular matrix with equal subdiagonals. Weight vectors with five nonzero weights were found to be accurate for most cases. More weights can be employed where indicated by appropriate diagnostics.

## Generalizability Issues

One issue involved in TSA is generalizability, that is, determining how the results from a single individual or unit can be generalized to a larger population. Although generalizability might not always be the primary goal for time-series studies, it can be explored fruitfully using the following methods.

One approach is *systematic replication,* which relies on logical inference rather than formal statistical inference. A weakness of this approach is that the judgment of the reviewer plays a critical role in the conclusions reached.

Two quantitative approaches have been developed that combine multiple replications of a time-series study. A *pooled time-series analysis* involves the direct combination of the data from different time-series studies. In this section, only one approach is described, an extension of the general transformation approach. This method uses a patterned transformation matrix, where a single vector represents observations of all the units. This vector contains the set of subvectors for the individual units combined in the form of a single vector. A properly parameterized design matrix permits comparisons among different units. The basic form of the design matrix should be based on the analyses of the individual units and/or a priori knowledge.

An alternative procedure to combining data from several individuals or units is *meta-analysis*. Procedures for performing a meta-analysis have been well developed for traditional experimental designs. However, meta-analysis procedures have not been widely applied to single-subject designs. A problem is developing definitions of effect size that are appropriate for time-series data.

## Multivariate Time-Series Analysis

A time-series analysis on a single dependent measure involves many procedures common to multivariate statistics because the following two vectors of unknowns must be estimated simultaneously: (1) the vector of parameters and (2) the vector of dependency coefficients. However, when assessing a single unit or subject on multiple occasions, two or more variables can be observed on each occasion. The term *multivariate time series* is used here to denote the observation of more than one variable at each point in time. The variables might be viewed conceptually as including both dependent and independent variables, or multiple dependent variables. If some of the observed variables are independent variables, then the appropriate analysis is the time-series equivalent of an analysis of covariance. In time-series designs, there are two unique problems. First, the investigator must determine the appropriate lag between the covariate(s) and the dependent variables. Second, there might be dependency present in the covariate(s), requiring a transformation before performing the analysis. One application is to control the effects of seasonality in the data (see below).

If the variables can be viewed as a set of dependent variables (i.e., multiple indicators of one or more constructs that form the outcome space of interest), then the appropriate analysis would be the time-series equivalent of a multivariate analysis, sometimes described as dynamic factor analysis. This method involves merging the longitudinal data approaches employed in TSA with the use of latent variables or factors to organize a set of observed variables. The recently developed dynamic factor analysis permits serial dependency in the data. Dynamic factor analysis might prove especially useful to behavioral researchers interested in questions of growth or change over time, as well as the underlying processes, because complex serial relationships among variables can be explored using this methodology. One limitation to implementing a dynamic factor analysis in practice is the number of observations required to provide an adequate sample estimate of the population covariance matrices. Other issues include determining the

correct number of factors to extract and determining the correct lags between the variables in the final model. However, the use of multiple indicators to measure one or more latent variables represents a promising means of extending the focus of TSA from univariate to multivariate outcome spaces.

## Cyclic Data

The presence of cyclic or seasonal data is another potential confounding variable in time-series data. Daily data gathered on individuals often have a weekly or monthly cycle. The following three alternative procedures have been proposed to deal with cyclic data: (a) *deseasonalization,* which assumes that the cyclic nature of the data is well known and the data can be adjusted for seasonal effects before it is reported, based on a priori information, prior to any TSA; (b) *statistical control,* which identifies a variable that is sensitive to the same seasonal effects as the dependent measure but cannot be affected by the intervention and thus used as a covariate; and (c) *combined models,* which involve two models, one that demonstrates the dependency and a second that demonstrates the seasonal component.

## Missing Data

The problem of missing data is almost unavoidable in TSA, and it presents several unique challenges. Life events will result in missing data even for the most conscientious researchers. Roderick Little and Donald Rubin provide the most thorough theoretical and mathematical coverage of handling missing data. In an extensive simulation study, Wayne Velicer and Suzanne Colby compared the following four different techniques of handling missing data in an ARIMA (1, 0, 0) model: (a) deletion of missing observations from the analysis, (b) substitution of the mean of the series, (c) substitution of the mean of the two adjacent observations, and (d) maximum likelihood estimation. Computer-generated time-series data were generated for 50 different conditions. The maximum likelihood procedure for handling missing data outperformed all others. Although this result was expected, the degree of accuracy was very impressive. The three ad hoc methods all produced inaccurate estimates for some parameters.

## Time as a Critical Variable

Another critical but often overlooked aspect of longitudinal designs is the importance of choosing an appropriate unit of time. TSA assumes that the observations are taken in equally spaced intervals. Unfortunately, very little information is available in the behavioral sciences to guide the choice of interval size. Sometimes the interval is predetermined, such as when existing data are employed. Other times, the choice of interval is determined by the convenience of the experimenter or the subject. As less obtrusive methods of data collection become available, and thus denser sampling of dependent variables are made possible, the choice of interval will be able to better reflect the needs of the research question.

The interval employed could strongly influence the accuracy of the conclusions that can be drawn. For example, in Figure 1, the choice of two 8-hour intervals (parts of the day when the subject was awake) strongly influenced the results. A negative autocorrelation of –.70 would become a positive autocorrelation of .49 if the observations had occurred once a day and .24 if assessed every 48 hours. In general, longer intervals can be expected to produce lower levels of dependency in the data.

If the focus is on the functional relationship between two variables, then the time interval can also be critical. If a change in $X$ produces a change in $Y$ with a 48-hour lag, then observations taken at weekly intervals might erroneously conclude that the variables are not related, whereas observations taken at 24-hour daily intervals would detect the relationship. Until there are adequate theoretical models and accumulated empirical findings for the variables of interest, shorter intervals will be preferable to longer intervals because it is always possible to collapse multiple observations. It is also important that any statements about the presence or degree of a relation between variables based on autocorrelations and cross-lagged correlations always reference the interval employed.

## Telemetrics

Recent advances in *telemetrics* (for a review see work by Matthew Goodwin, Wayne Velicer, and Stephen Intille)—a class of wireless information systems technology that can collect and transmit

a wide variety of behavioral and environmental data remotely—permit an unprecedented amount of longitudinal data to be gathered for time-series studies. Telemetrics include "wearable computers" that weave on-body sensors into articles of clothing, "ubiquitous computers" that embed sensors and transmitters seamlessly into the environment, and handheld devices such as mobile phones and personal digital assistants that can record cognitive and affective states. With this technology, the input or signal is repeatedly assessed, quantified, and combined with timing data, resulting in a single data stream for researchers to explore.

### Conclusion

Time-series analysis and telemetric monitoring have tremendous potential for enhancing behavioral science research. TSA is one of the many computational procedures that have been developed specifically for the analysis of longitudinal data since the 1970s. In fact, TSA can be viewed as the prototypic longitudinal method. The combination of computational advances and new sources of data has increased the number of potential applications. We are now reaching the point in the behavioral sciences where the data analysis method will be matched to the research problem rather than the research problems being determined by the available methods of data analysis.

*Wayne F. Velicer, Bettina B. Hoeppner, and Matthew S. Goodwin*

***See also*** Autocorrelation; General Linear Model; Least Squares, Methods of; Longitudinal Design; Meta-Analysis; Multiple Regression

#### Further Readings

Algina, J., & Swaminathan, H. A. (1979). Alternatives to Simonton's analysis of the interrupted and multiple-group time series designs. *Psychological Bulletin, 86,* 919–926.

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (1994). *Time series analysis: Forecasting and control* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Glass, G. V., Willson, V. L., & Gottman, J. M. (1975). *Design and analysis of time series experiments.* Boulder, CO: Colorado Associate University Press.

Goodwin, M. S., Velicer, W. F., & Intille, S. S. (2008). Telemetric monitoring in the behavior sciences. *Behavior Research Methods, 40,* 328–341.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: Wiley.

Molenaar, P. C. M. (1985). A dynamic factor model for the analysis of multivariate time series. *Psychometrika, 50,* 181–202.

Molenaar, P. C. M. (2008). On the implications of the classical ergodic theorems: Analysis of developmental processes has to focus on intra-individual variation. *Developmental Psychobiology, 50,* 60–69.

Simonton, D. K. (1977). Cross-sectional time-series experiments: Some suggested statistical analyses. *Psychological Bulletin, 84,* 489–502.

Velicer, W. F., & Colby, S. M. (2005). A comparison of missing data procedures for ARIMA time series analysis. *Educational and Psychological Measurement, 65,* 596–615.

Velicer, W. F., & Fava, J. L. (2003). Time series analysis. In J. Schinka & W. F. Velicer (Eds.), *Research methods in psychology* (pp. 581–606). New York: Wiley.

Velicer, W. F., & McDonald, R. P. (1984). Time series analysis without model identification. *Multivariate Behavioral Research, 19,* 33–47.

## Time Studies

The entity or factor of time is an omnibus concept that has served well the enterprises of experimental science, research design, and scientific analysis in the physical, biological, behavioral, and social sciences. Indeed, time itself is often the object of study in the sciences, particularly in physics and experimental psychology. In essence, the concept of time seems to be so abstract, as well as so fundamental, that no universally accepted definition across, and within, the sciences is available. Thus, it is an issue of debate whether time is an abstraction dealing with the marking of change or a medium through which change occurs including endogenous and/or exogenous "clocks" or "timers." In physics, whereas Newtonian mechanics regarded time as absolute, the newer notion of the special theory of relativity asserts that time is relative to motion. Also, many theories in philosophy and psychology posit differences in the absolute versus relative nature of time and temporal experience. This entry discusses various definitions

and designs, measurements and methods, and the future direction of time studies.

## Definitions and Designs

The area of psychological research investigating time and temporal dimensions might be subdivided into other traditional rubrics, such as time-sense (apprehension of duration, change, or order of occurrence), time estimation and perception (attention to, or apprehension of, change through the integration of a series of stimuli and characterized by the ability to conceive of duration, simultaneity, and succession), and time orientation, horizon, or perspective (the totality of one's views of his or her psychological future and past existing at a given time, and is reflective of one's distinctive position in society and one's unique developmental history), among other designations. Additionally, time has served as a basis for several important design strategies in research. For example, the quasi-experimental design called time series is a *within-subjects* design where the performance of a single group of participants is measured both before and after the experimental treatment is administered; also, in the time design approach called *single-group pretest–posttest design,* a temporal comparison is made between a single pretest measure or score and a single posttest measure.

Other time-related designs include the *interrupted time-series design* (employment of several pretest and several posttest measurements to control for normal or natural fluctuation rates), the *multiple time-series design* (a between-subjects design involving the use of a control group of participants in addition to the experimental participants), the *retrospective and ex post facto designs* (the use of empirical procedures to suggest meaningful relationships between events that have occurred in the past), the *repeated measures design* (an application of the within-subjects design that uses repeated measures as one factor in a factorial design), and the *mixed designs* (a common design used in psychological research that combines a between-subjects design with a within-subjects design, and involves complicated statistical analyses including computations of error terms and F-ratios).

In the case of the concept of time perspective in particular, several research strategies might be suggested: the length of time in which thought is projected ahead (protension), the length of time in which thought is projected back into the past (retrotension), the number of events thought about in past or future (density), the degree of organization within one's past–present–future matrix (cohesiveness), and the sense of perceived rate of movement toward the future (directionality). Moreover, the basic distinction between retrospective and prospective time is reflected in research designs, most notably in retrospective versus prospective designs and studies. In the former, participants with a particular characteristic or disorder are studied and their histories are examined to determine the relevant causal factors and relationships; in the latter, a cohort of participants is studied currently over a period of time to determine factors that might cause a particular characteristic or disorder to emerge. Typically, the findings of retrospective studies carry less weight (that is, they are more open to alternative explanations of results, alternative interpretations, or rival hypotheses) than those of prospective studies largely because they rely more on the fallible memories of the participants. In general, in *retrospective time* designs, it has been found that when time is filled with varied and interesting experiences, it seems short in passing but long as one retrospectively looks back (in contrast, a tract of time empty of experiences seems long in passing but short in retrospect), whereas in *prospective time* designs, findings indicate that tracts of time seem to be shortened in passing whenever one is so fully occupied with their content as not to note the actual time itself (thus, a day full of excitement with no pause is experienced as passing quickly, but a day full of waiting seems to pass slowly).

In each of these designs, time might be treated as an independent (stimulus, predictor, treatment, or cause) variable or as a dependent (response or effect) variable, whereas in the former, it is the variable in an experiment that is manipulated specifically or observed to occur before the dependent variable and that might, or might not, be related causally to the dependent variable. In the area of psychological and subjective time research, the following independent variables have been incorporated successfully into research designs: marker and boundary stimuli; filled versus empty intervals; hypnosis and dreams; interpolated stimuli;

isolation, monotony, and satiation; motivation, goals, and needs; physiological factors; pitch and sound intensity; practice and task difficulty; retardation and pathology, rhythmical intervals; sensory modalities; reinforcement schedules; spatial factors; and temporal order and anchors. Also, the following dependent variables have been identified and/or employed successfully in time research designs: the time-order error (the tendency to overestimate the second stimulus in relation to the first in a situation where one stimulus follows another in time), temporal discrimination, brightness and pitch discrimination, duration estimation, temporal indifference interval, time acceleration, time comprehension, time compression, time distortion, and duration judgments in relation to distance and speed judgments.

In psychology, temporal experiences of individuals have been examined mainly via *single-subject* (formerly called case studies) or *small-N* designs where—in the tradition of Ivan Pavlov, Hermann Ebbinghaus, Edward L. Thorndike, and Burrhus F. Skinner—issues in the areas of perception, conditioning, learning, memory, and psychophysics have been studied. Today, the single-subject designs might involve more than one participant in the research and be referred to as small-*N* designs; in such research strategies, the individual's data are not averaged with those of other participants but emphasize the intensive experimental examination of processes within the individual participant.

Various dimensions of temporal experience have been paired uniquely with various research designs. For example, the notions of time perspective, temporal orientation, and time horizon lend themselves well to the *case study designs* (a narrative description of an individual or some aspect of an individual that combines relevant features of the person's history and current situation), whereas the notions of time estimation and time perception are amenable to the *factorial designs* (the use of two or more independent and dependent variables or factors in a single experiment).

## Measurement and Methods

In the context of scales of measurement in research design, time might be embedded in one, or more, of the traditional four measurement scales: nominal (e.g., verbal labels of early childhood versus late adulthood), ordinal (e.g., first, second, and third place in a foot-race), interval (e.g., hours of the day or days on a calendar), and ratio (e.g., a stop-watch, clock, or other time-measurement device used to measure physical and cognitive reactions). In the early experimental studies of time sense and time perception, circa 1894, it was customary to use a time-sense apparatus—an instrument for measuring the accuracy of time estimations that consisted of a uniformly rotating metal arm that during rotation came into contact with two or more sets of terminals or contacts, producing similar sounds or other forms of stimulation separated by a time interval that depended on the rate of rotation and the distance between the contacts. Today, of course, temporal events in the laboratory are administered and measured by high-speed and precise computer clocks and timers.

The time dimension has been studied by a wide array of methods, and, in psychology in particular, the recruitment of the psychophysical approach to the issue has yielded a so-called psychophysics of duration in which time perception recently has been partitioned into subcategories of duration scaling and duration discrimination. In the duration scaling method, the researcher examines participants' perceived durations of temporal intervals, and in the duration discrimination method, the analysis focuses on the variability of a group of responses as participants attempt to discriminate among relatively short temporal intervals that might be confused. Furthermore, within the method of duration discrimination, assessments of judging temporal intervals have included other dimensions, such as the effects of additional variables (e.g., interval markers across several sensory systems) on perceived duration or on the validity and variability of the Weber law or fraction vis-à-vis temporal experience experiments on both humans and animals.

## Future Directions

Future directions in time study include the examination of interdisciplinary differences in temporal experience, investigations in the physical sciences involving areas such as expanding versus contracting universes, cyclical versus linear time, entropy

increases, arrows of time, time travel, time atoms, cosmic time, the beginning and end of time, and relativistic versus absolute time. Future investigations on temporal experience in psychology will include areas such as social and personality processes, gender issues and differences, developmental stages, motivational and memory dynamics, discrimination processes, micro-waiting behaviors, environmental and stimulus features, cognitive and perceptual processes, consumer behavior, psychological disorders, and greater sophistication of theoretical models concerning internal and biological clocks and mechanisms, as well as improved empirical and experimental designs for studying time.

*Jon E. Roeckelein*

***See also*** Dependent Variable; Factorial Design; *F* Test; Independent Variable; Levels of Measurement; Psychometrics; Repeated Measures Design; Single-Subject Design; Time Series Study; Weber–Fechner Law; Within-Subjects Design

**Further Readings**

Grondin, S. (2008). *The psychology of time*. Amsterdam, the Netherlands: Elsevier.
Mitchell, M., & Jolley, J. (2001). *Research design explained*. New York: Harcourt.
Ray, W. J. (2003). *Methods toward a science of behavior and experience*. Belmont, CA: Wadsworth.
Roeckelein, J. E. (2000). *The concept of time in psychology*. Westport, CT: Greenwood.

# Treatment(s)

The word *treatment* appears many times in the typical text on statistics and/or research design. It appears frequently in the indices of such texts. It is rarely defined. It is defined here, and how the term is used by researchers is shown.

## Treatments, Treatment Effects, Independent Variables, and Experimental Research

Researchers are most likely to use the word *treatment* when referring to experimental research, especially when the data from that research were analyzed via analysis of variance (ANOVA). In experimental research, the researcher manipulates the independent or treatment variable(s) and then observes whether the treatment groups differ on one or more dependent or outcome variables.

## Multiple-Case Research

In multiple-case research, the scores of two or more groups of cases (which might be the same research units or might be different research units) are compared to determine whether a treatment effect exists.

### *Two-Treatment Research*

Consider a research study designed to determine the effects of damage to a particular nucleus in the brain. The researcher randomly assigns 20 rats to each of two groups. The rats in the one group have an electrode placed at the location of the nucleus of interest, and then electrical current is applied to damage that nucleus. The experimental treatment here involves everything done to these rats as part of the research—how they are housed, fed, and prepared for surgery; the placement of the electrode in the brain; the electrolytic damage to the nucleus; and so on. The other group of rats receives a different treatment—sham surgery. They are treated exactly the same as the lesioned group, with the exception that the electrical current is not applied to the electrode that is placed within the brain. Sometimes such a group is called a control group. For this group, the (control) treatment is how they are housed, fed, and prepared for surgery; the placement of the electrode in the brain; and so on, but no electrolytic damage is done to the nucleus. It is important that the two groups be equated on all aspects of their treatment with the exception of the one (or more) aspect(s) of interest to the researcher. Accordingly, when the two groups of rats are compared on some outcome variable(s) of interest, any significant differences found can be attributed to the treatment variable.

For this example, the treatment variable is "type of surgery," and it has two values, "lesion surgery" and "sham surgery." After the animals have recovered from the surgery, they are tested on one or more outcome variables and then the groups are compared on the outcome variable(s). If the

outcome variable is categorical (for example, does the rat approach or flee when a strange conspecific is presented), then the data might lend itself to a contingency table analysis (typically done with chi-square), which, if statistically significant, will lead to the conclusion that the treatment variable does affect the outcome variable. If the data are continuous and normally distributed, then a *t* test or ANOVA would typically be employed to determine whether the treatment variable affected the outcome variable. If the outcome variable is continuous but not normally distributed, then one would typically employ one of the several available techniques that do not require normality.

As noted earlier, the term *treatment* is most likely to be used when the data were analyzed with an ANOVA. For the research just described, suppose that the outcome (dependent) variable was the number of trials it took the rat to learn some new task and that the data were normally distributed. The ANOVA would partition the sums of squares (precursor to variance) for the outcome variable into two sources. One source is the treatment sum of squares, which is also known as the between-groups sum of squares or the model sum of squares. The other source is the within-group, error, or residual sum of squares. Each of these sums of squares is divided by its degrees of freedom to obtain a mean square. The treatment mean square estimates the variance as a result of the treatment and error, whereas the error mean square estimates the variance caused by error alone. The ratio of the treatment mean square to the error mean square is the *F* statistic, which is used to obtain the conditional probability of getting sample groups that differ from one another by at least as much as those obtained if the groups did not differ in the populations from which they were sampled. In other words, the *F* and the *p* values obtained in an ANOVA evaluate the null hypothesis that there is no treatment effect.

### *ANOVA and Nonexperimental Research*

The ANOVA might also be employed to analyze data from nonexperimental research. For example, female and male clients of a medical clinic are asked how satisfied they are with the services received at the clinic. The satisfaction scores are normally distributed. ANOVA is used to compare the two groups, which consist of men and women. The between-groups sum of squares here would not appropriately be called a treatment sum of squares, because the researcher did not (presumably) treat the two groups differently. Most researchers would not refer to the gender variable as a treatment variable, although some might refer to it as an independent variable, which can be confusing. With nonexperimental research like this, some researchers believe that if they conduct an ANOVA and call gender the "independent variable," then significant results allow them to make a causal inference, but if they perform a (mathematically equivalent) correlation/regression analysis, then they cannot make a causal inference.

In some cases, it might be appropriate to employ the term *treatment* even with nonexperimental data. Suppose that patients at the clinic with a particular medical condition are treated with either one drug or another. They are not randomly assigned to groups. Which drug is prescribed depends in part on the preference of their physician and in large part on the formulary used by the provider of their medical insurance. To compare the two groups on posttreatment severity of symptoms, again, assuming normality of the outcome variable, ANOVA could be employed. Because the two groups were treated differently (with respect to which medication they received), it would be reasonable to call the between-groups sum of squares a treatment sum of squares and to call the difference between the groups a treatment effect. An interpretation of the results in this case would, however, need to be tempered by recognition that the two groups might well have been nonequivalent prior to being treated with one or the other drug.

### *More Than Two Treatments and More Than Two Treatment Variables*

An experiment might involve more than two treatments. For example, one could add to the lesion surgery versus sham surgery experiment a third group of rats that received no surgery (sham or lesion) at all. In this case the treatment variable has three levels—no surgery, sham surgery, and lesion surgery.

An experiment might involve more than two treatment variables. For example, one could add to the lesion research a second treatment variable, the dose of an experimental drug administered to the rats after the surgery. To investigate the effects of an experimental drug that is hypothesized to help animals recover from a brain lesion, rats might be assigned randomly to the following three treatment groups: one receiving only a placebo injection, a second receiving an injection with a low dose of the experimental drug, and a third receiving an injection with a high dose of the drug. Such a layout is referred to as a factorial design. In this case, it is a $3 \times 3$ factorial design. One factor or treatment is the type of surgery with three levels (none, sham, or lesion), and the other is the dose of the drug (none, low, or high). The factorial combination of the two trichotomous independent variables produces $3 \times 3 = 9$ treatment combinations (often called "cells"): No surgery and no drug; no surgery and low dose of drug; ... lesion surgery and high dose of drug. The factorial analysis would test four treatment effects, as follows: the effect of the combination of the type of surgery and the dose of drug, the effect of the type of surgery ignoring the dose of the drug, the effect of the dose of the drug ignoring surgery, and the interaction between the type of surgery and the dose of drug.

A factorial experiment might include both an experimental factor and a nonexperimental factor. One example is the randomized blocks design. The experimental units (subjects or cases) are matched up into blocks of $k$ (where $k$ is the number of levels of the treatment variable) such that within each block, the subjects are identical or similar on the matching variable(s). In some cases, the subjects within a block are identical in all aspects; that is, they are the same person, animal, or other thing. In this case, the treatment variable is referred to as within subjects or repeated measures, and each subject is exposed to each of the $k$ treatments. The ANOVA used with such a design is sometimes called a treatment $\times$ block ANOVA. The blocking variable is treated as a second factor in a factorial design. The analysis partitions the total variance into three sources: treatment, blocks, and treatment $\times$ blocks. The treatment $\times$ blocks mean square is employed as the error term for the $F$ used to test the effect of the treatment variable. This analysis removes, from what otherwise would be error variance, the effects of the variable(s) on which subjects were matched when creating the blocks. This reduction in error variance should increase power, the probability that the analysis will detect as significant the effect of the treatment variable (assuming it has an effect). This increase in power depends on whether the researcher chose matching variables on which the subjects differ and that are well correlated with the dependent variable.

## Single-Case Research

The term *treatment* is also employed in single-case research. One common design used in single-case research is the A-B-A withdrawal design. In this design, "B" stands for the experimental treatment and "A" for the control treatment. For example, a researcher might be interested in the effects of a particular behavioral contingency management program for a child with a developmental disability. In phase $A_1$ (baseline) of the design, the target behavior(s) [dependent variable(s)], such as frequency of various behaviors, are measured until relatively stable across time. During phase B, the experimental treatment (the behavioral contingencies and the intervention) is introduced, and it is observed whether the dependent variable(s) change. Phase $A_2$ involves withdrawal of the experimental treatment and continued observation of the dependent variable(s). The treatment might be implemented a second time (ABAB) or an alternation of baseline and treatment might occur several times. If the treatment effect is observed every time, the treatment is introduced and the behaviors return to baseline every time the treatment is removed. Then, the researcher can be relatively confident that the change in behavior accompanying the treatment is caused by the treatment rather than by factors such as maturation, history, and other threats to internal validity.

*Karl L. Wuensch*

***See also*** Analysis of Variance (ANOVA); Block Design; Control Group; Dependent Variable; Experimental Design; Factorial Design; Independent Variable; Internal Validity; Single-Subject Design; Sums of Squares

**Further Readings**

Cooper, J. O., Heron, T. E., & Heward, W. L. (2007). *Applied behavior analysis* (2nd ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Kazdin, A. E. (1982). *Single-case research designs: Methods for clinical and applied settings*. New York: Oxford University Press.

Kirk, R. E. (1995). Research strategies and the control of nuisance variables. In *Experimental design: Procedures for the behavioral sciences* (3rd ed.) (pp. 1–28). Pacific Grove, CA: Brooks/Cole.

Mendenhall, W., & Beaver, R. J. (1991). Experimental design and the analysis of variance. In *Introduction to probability and statistics* (8th ed.) (pp. 529–580). Boston: PWS-Kent.

Mook, D. G. (1982). The analytic experiment. In *Psychological research: Strategy and tactics*. New York: Harper & Row.

Wuensch, K. L. (2007). *Independent variables and dependent variables*. Retrieved May 7, 2008, from http://core.ecu.edu/psyc/wuenschk/StatHelp/IV-DV.htm

# Trend Analysis

Trend analysis is the assessment of linear and curvilinear associations among variables. Linear trends are straight-line associations commonly evaluated in ordinary least squares (OLS) regression. Curvilinear trends include U- or inverted U-shape associations known as quadratic or parabolic trends, as well as sideways S-shape associations known as cubic trends. Other types of trends are available but rarely pursued because justifications for their use are often questionable. After describing some applied examples and the background of trend analysis, this entry discusses how trend analysis is applied in research.

## Examples and Background

Applied examples of trend analysis within the research literature are prevalent. For example, Debra Murphy and William Marelich report linear trend declines in depression during an 18-month period for both resilient and nonresilient children whose mothers are HIV positive. Jessica Kahn and colleagues investigated patterns of adolescents' physical activity, noting a quadratic trend as physical activity peaks at the age of 13 then declines (an inverted U-shape). Adele Hayes and her associates found that the effects of exposure-based cognitive therapy (EBCT) on depression during the course of therapy could best be summarized as a cubic trend, with depression initially declining after treatment, followed by an increase or depression spike, then another decline in depression (a sideways S-shape).

Trend analysis and assessment of curvilinear associations emerged because of the limitations in assuming variable relationships are linear. In the latter two examples noted earlier, assessing a linear trend alone would not have been sufficient to explain the variable associations. To illustrate, say an investigation was performed examining the association between task performance and arousal level. Using the Yerkes-Dodson law as the theoretical framework, performance and arousal should have an inverted U-shape; moderate levels of arousal should lead to optimal performance, whereas low and high levels of arousal should show arrested performance. A bivariate scatter plot of hypothetical task performance and arousal data reveals the expected curvilinear association. However, a bivariate OLS regression between these two variables would not capture this inverted U-shape association because only the linear trend is being assessed, and the multiple R value would be close to zero. The variables are related to each other but in a curvilinear fashion. To reveal this relationship, a quadratic trend should be assessed to summarize the inverted U-shape association between the two variables.

## Trend Analysis in Application

Trend analysis is commonly performed in OLS multiple regression and analysis of variance (ANOVA) applications. It is also used in statistical applications designed to assess associations such as structural equation modeling and multilevel modeling. Trend analysis is performed when there is a hypothesized linear and/or curvilinear association between two or more variables, with one of the variables acting as a criterion or dependent variable and the others acting as predictors. In addition, the predictors must be scale measured to allow the shape of the association to have meaning. For example, a predictor might be

a continuous score on a measured psychological scale, such as a score on the depression inventory. Or, the predictor might be a discrete measure such as group membership, with each group representing participants receiving different doses of a drug or possibly each group representing additive intensity levels of an independent variable. As long as the discrete variable is scale measured, it might be used for trend analysis.

Trend analysis cannot be used if the predictor is nominal because the resulting linear and curvilinear associations will have little meaning because of the arbitrary order of the discrete values. For example, a bivariate scatter plot between a nominal predictor variable such as color preference (e.g., 1 = white, 2 = blue, and 3 = red) and a scale measured criterion such as depression might show a curvilinear association, but such an association would be artificial because the values assigned for color preference are arbitrary. The values assigned to each color could have easily been 1 = blue, 2 = red, and 3 = white. Dummy coding a nominal variable such as color preference would assist in correctly evaluating the linear association of each color on depression, but investigations into higher order trends cannot be made because each dummy coded variable has only two levels. Therefore, an assessment beyond linear trends for nominal level predictors (assuming dummy coding is applied) cannot be made.

The two ways to apply trend analysis include powered polynomials and orthogonal polynomials. The term *polynomial* is a label for a mathematic expression applied to one or more variables. For trend analysis, polynomials are created by multiplying the predictor variables by specific values. Regardless of whether powered polynomials or orthogonal polynomials are adopted, both will produce the desired trends. As a rough rule of thumb, powered polynomials are used when the predictor variables are scale measured and continuous. When the predictor variables are scale measured and discrete, and represent a small number of scaled categories or groups, then orthogonal polynomials are often used.

As with most quantitative methods, trend analysis is performed with Occam's razor in mind. The linear trend must always be assessed first, followed by the higher order trends (i.e., quadratic, cubic, etc.). Fitting the more complex trends first can be problematic, especially when the terms are not orthogonal because the linear trend is evident in the higher order trends. Even if the trends are orthogonal, accounting for the simplest trend first retains the simplest explanation for the variable associations.

### *Powered Polynomials*

Powered polynomials refer to taking a predictor variable to a particular power to form the trends. The original variable ($x$) acts as the linear trend. The quadratic trend is formed by squaring the predictor ($x^2$). The cubic trend is formed by cubing the predictor ($x^3$). Additional trends are created in a similar manner using higher order powers. The powered polynomial approach is used when predictor variables are scale continuous, but it might also be used when the predictor variables are scale discrete.

Prior to creating powered polynomials, it is recommended that the predictor variable is mean centered. Mean centering the predictor prior to creating the powered polynomials will reduce problems associated with multicollinearity between the linear and quadratic trend. Because a predictor variable is being squared, the new variable will have a large correlation with the original predictor (e.g., greater than .98 for the original predictor and the newly formed quadratic trend). In some instances, such multicollinearity might cause computational problems. After mean centering, the correlation will be zero. However, centering will only relieve multicollinearity between the linear and quadratic trends. It will not relieve multicollinearity between the linear and other higher order trends.

Mean-centering is completed by taking each score on the predictor and subtracting it from the mean value of the predictor. This will produce a distribution with a mean of zero and will retain the original variable's standard deviation. An alternative is to center the variable through standardization. This will also produce a mean of zero but will change the standard deviation to a value of one. Either approach is acceptable. Once the predictor is centered, the powered polynomial is produced using the centered predictor. The centered predictor acts as the linear trend, whereas the square of the centered predictor is the quadratic

trend. Centering the data will not alter the overall assessment of variance accounted for in the criterion or significance of the trends being assessed. It will, however, have an effect on the final parameter estimates and their significance. If the OLS regression parameters and their significance are of import to the research, then centering the predictor is required.

After the powered polynomials are created, OLS multiple regression might be used to evaluate the different trends. This is done using a hierarchical approach, entering first the linear trend followed by the quadratic trend. If a cubic trend is of interest, then it is entered after the linear and quadratic trends. The hierarchical approach is used assuming the trend lines based on powered polynomials are correlated. Each trend (linear, quadratic, etc.) is entered into the regression model one by one, and it is evaluated at each step through the change in the model multiple *R*-square value. As this process unfolds, trend analysis can determine (a) whether the linear trend accounts for a significant amount of variance in the criterion and (b) whether the quadratic trend accounts for a significant amount of variance in the criterion, above and beyond that of the linear effect. These analyses continue as higher order trends are added to the model. For each of the steps in the hierarchical analysis, the change in the model multiple *R*-square value and its significance are examined. In addition, assuming the powered polynomial for the higher order trends were created from a centered predictor, the subsequent regression parameter estimates and their significance values might be used for subsequent regression applications.

### *Orthogonal Polynomials*

Orthogonal polynomials for trend analysis are created by taking a scale discrete predictor variable and multiplying the category or group means by a set of contrast weights. The resulting products are used to assess the trends. Orthogonal polynomials are commonly used in ANOVA applications and are specific contrasts constructed to produce the trends of interest. They might also be used in OLS regression or similar applications.

The orthogonal polynomials might be derived by the researcher or found in the appendices of most advanced statistical textbooks. In addition, statistical programs such as SPSS®, an IBM company, have options that produce orthogonal polynomial results (IBM® SPSS® Statistics was formerly called PASW® Statistics). Unlike powered polynomials, the predictors do not need to be centered. However, one limitation for orthogonal polynomials is that the number of individuals within each category of the predictor should be equal. If the numbers within each scaled category or group are unequal, then the resulting trends will not be orthogonal, and additional steps must be taken to ensure the accuracy of the results.

When using orthogonal polynomials, the first step is to identify the number of scaled categories or groups in the predictor variable. The number of trends that can be fit is equal to the number of categories in the predictor variable minus one ($k - 1$; $k$ being the number of categories or groups) and follows the typical trend sequence (linear, quadratic, cubic, etc.). If the predictor has three categories, then only the linear and quadratic trends might be fit. If there are four categories, then three orthogonal polynomials might be fit.

The easiest way to derive the orthogonal polynomials for trend analysis is to refer to an advanced statistics textbook, which has an appendix of orthogonal polynomial mean contrasts. For a three-category predictor variable, a standard textbook appendix would show two orthogonal polynomial contrasts: {–1 0 1} for the linear trend, and {1 –2 1} for the quadratic trend. Look closely at these mean contrast values. The contrast applied to the means for a linear trend follows a linear trajectory, whereas the contrast for a quadratic trend is a U-shape. If the predictor variable had four levels, then the three orthogonal polynomial contrasts are {–3 –1 1 3} for the linear trend, {1 –1 –1 1} for the quadratic, and {–1 3 –3 1} for the cubic trend. Assuming equal category or group sizes, these mean contrasts are independent of each other. Ways to proof the independence of the contrasts might be found in most advanced textbooks.

Once the orthogonal polynomials have been applied, they might be used in the typical manner for assessing contrasts. Assuming equal category or group sizes, the orthogonal polynomials will be independent, and their effects might be evaluated simultaneously if ANOVA is the statistical application of choice. However, if unequal category sizes are evident, then the resulting trends will not be

independent, and a hierarchical approach should be adopted. Interpreting trend analysis results using orthogonal polynomials is done in the same fashion as with powered polynomials. The linear trend takes precedent, followed by the quadratic and higher order polynomials.

*William David Marelich*

***See also*** Analysis of Variance (ANOVA); Least Squares, Methods of; Multiple Regression; Occam's Razor; Orthogonal Comparisons; Scatterplot; U-Shaped Curve

**Further Readings**

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Howell, D. C. (2002). *Statistical methods for psychology* (5th ed.). Pacific Grove, CA: Duxbury.

Keppel, G. (1982). *Design and analysis: A researcher's handbook* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Keppel, G., & Zedeck, S. (1989). *Data analysis for research designs: Analysis of variance and multiple regression/correlation approaches.* New York: Freeman.

# TRIANGULATION

The term *triangulation* refers to the practice of using multiple sources of data or multiple approaches to analyzing data to enhance the credibility of a research study. Originating in navigational and surveying contexts, triangulation aligns multiple perspectives and leads to a more comprehensive understanding of the phenomenon of interest. Researchers differ in the emphasis placed on the purposes of triangulation; some investigators view it as critical to establishing corroborating evidence, and others focus on its potential to provide multiple lines of sight and multiple contexts to enrich the understanding of a research question. Particularly associated with qualitative research methods, triangulation typically involves examining data from interviews, focus groups, written archives, or other sources. Triangulation is often used in studies that combine both quantitative and qualitative approaches, and it is sometimes referred to as mixed methods or multimethod research.

## Types

Norman Denzin identified four types of triangulation. First, *data triangulation* involves using multiple sources of data in an investigation. In a research study examining hospital staff morale, for example, interviews with medical personnel might be compared and cross-checked with staff surveys and records of focus groups consisting of hospital employees. Second, *investigator triangulation* involves employing several evaluators to engage in observations or analyze participant responses. Using multiple investigators allows for the auditing of data consistency and reduces the potential bias inherent in employing only one investigator or analyst. For example, a group of researchers analyzing responses to open-ended survey questions might be less likely to draw erroneous conclusions than a single investigator, whose expectations might color interpretations of the data. A related practice known as *member checking* involves having study participants review transcripts and the findings derived by investigators to verify the accuracy of their recorded responses and comment on the conclusions drawn. Third, in *theory triangulation,* multiple theoretical perspectives are considered either in conducting the research or in interpreting the data. Employing a multidisciplinary team is one approach that brings different theoretical perspectives to bear on the research question. Last, *methodological triangulation,* which is the most commonly used form of triangulation, engages multiple methods to study a single problem. Typically employed to compare data collected through qualitative methods with quantitative data, methodological triangulation can establish the degree of compatibility between information obtained through different strategies. Qualitative and quantitative methods might be employed simultaneously (e.g., distributing a questionnaire and conducting a case study) or might be used in a sequential fashion (e.g., a pilot study serves as the foundation for a randomized controlled trial conducted at a later date). Methodological triangulation might take the form of within-methods triangulation, where multiple

quantitative or qualitative approaches are employed, or between-methods triangulation, where both quantitative and qualitative approaches are used. Within-methods triangulation, on the one hand, has been criticized as a weaker strategy, as it only employs one method (either qualitative or quantitative) and does not compensate for the limitations of the particular paradigm. Between-methods triangulation, on the other hand, offers the possibility that the biases inherent in one approach will be mitigated by the inclusion of other sources of data, methods, and investigators.

Studies that employ triangulation typically yield one of three outcomes: convergence, inconsistency, or contradiction. The particular outcome of the study challenges the researcher to bring together theory, research, and practice in the construction of a comprehensive explanation of the results.

## Debates Regarding the Purposes and Contributions of Triangulation

Researchers have disagreed regarding the purposes and potential contributions of triangulation. These differences emerge out of the paradigms underlying investigators' approaches to research. Researchers influenced by the positivist or postpositivist philosophies have viewed triangulation primarily as a means of overcoming the limitations inherent in using only one approach to research. They have viewed the benefits of employing multiple data sources as a means of verifying the findings of different methods, asserting that if data from two or more sources converge on the same information, the likelihood of error is reduced. This viewpoint is closely akin to the traditional criteria of reliability and validity in quantitative approaches. If a researcher finds that data from multiple sources corroborate in support of the same conclusion, then the researcher can be more confident in its validity.

In contrast, researchers influenced by a constructivist philosophy consider the benefits of triangulation to lie not in its potential to verify information but in its capacity to provide multiple viewpoints on the phenomenon of interest and to amplify the perspectives of participants who have been ignored or overlooked in traditional scientific inquiry. To constructivists, triangulation offers the opportunity to deepen the understanding of the research question and to explore multiple realities. Rather than viewing participant checking as validation or verification, for example, constructivists conceptualize it as obtaining additional data to expand the understanding of the research problem. To constructivists, triangulation is less concerned with attending to converging data than attending to obtaining multiple perspectives.

Some constructivists caution that researchers should avoid drawing links between qualitative research and quantitative methods as it might be construed as an effort to legitimize qualitative research to traditional quantitative audiences. They argue that by framing triangulation as a means of establishing validity or reliability, constructivists take an apologetic stance in trying to fit their methods into a traditional paradigm that postpositivists will appreciate and accept. This practice perpetuates the belief that qualitative research is not as rigorous or legitimate as "real science."

## Benefits

The benefits of triangulation vary depending on the perspective of the researcher. Postpositivists assert that triangulation enables researchers to minimize the biases inherent in using a single research approach. Studies that employ several methods and yield multiple types of data provide the opportunity for comparing and cross-checking findings. Because every type of data has strengths and limitations, using a combination of techniques helps compensate for the weaknesses found in one approach. Observations, for example, are limited in that the observer might focus attention on one particular aspect of the situation, while overlooking other, more significant events. The presence of the observing researcher might impact the participants in undetected ways. Likewise, it is difficult to determine whether the behaviors observed are typical of the participants or represent a limited snapshot of an uncommon action. Interviews might yield data that are affected by the interviewer's unique style of communication, by the personal recall or interpretation of responses, or by the interviewees' assumptions about, or reactions to, the interviewer. Archived documents tend

to be limited by the specificity of the information contained in written records, biases of the document writer, or distortion of information. Because of the inherent limitations of these and other research approaches, investigators can enhance the credibility of findings by building in the use of multiple sources of data through triangulation. This strategy enables the researcher to capitalize on the strengths of each approach and reduce the impact of the weaknesses inherent in a single approach.

Constructivist researchers, in contrast, assert triangulation is beneficial in that it allows the airing of multiple perspectives on the problem and prompts the researcher to consider multiple realities. They argue that traditional approaches to scientific inquiry have largely overlooked individuals who have had less power and influence. Triangulation invites members of these groups to have a voice in determining reality and in contributing to the expansion and proliferation of knowledge.

## Limitations

Although triangulation offers many benefits to the study of complex phenomena, several limitations have been identified. From a postpositivist perspective, triangulation does not always reduce bias. A researcher might triangulate using data collected through different methods (e.g., self-reports, diaries, and political speeches), but if that data come from a common source, such as a single person, then bias remains. Even if findings are corroborated from two different sources, a researcher cannot guarantee that both sources do not yield data that are flawed. A researcher's conclusions based on this information inevitably will be impacted. Thus, using triangulation to strengthen a study's validity is not always an effective strategy.

In studies that use methodological triangulation, it is not uncommon to obtain conflicting data. Investigators disagree regarding how to interpret discrepant findings. Constructivists argue that one should not necessarily expect triangulation to lead a researcher to a single truth, as there are multiple constructions of truth. Triangulation, they assert, should be viewed as a tool to enrich the process of inquiry and to allow multiple perspectives to emerge. Others note that discrepancies indicate different methodologies have captured different information. The skilled researcher is then challenged to make sense of the differences.

Some researchers have cautioned that different research methodologies should not be combined without a clear rationale. Data derived from different approaches cannot necessarily be compared and viewed as equivalent in their ability to address a research question. For example, data collected from journal entries might differ dramatically from data collected in interviews, because one approach elicits private thoughts, whereas the other taps communication within a social context. When using methods triangulation, researchers should be able to articulate a clear rationale for combining different methods.

Triangulation can be expensive and time consuming. Most research projects are limited by financial and time constraints. Therefore, using multiple investigators, measures, theoretical perspectives, and methods will depend on the resources allocated for the study. The potential benefits of triangulation must be weighed against the practicalities of budget and time frame.

An additional consideration in using triangulation involves the experience level of the primary investigator and the research team. Because studies employing triangulation can be complex, it is critical that the researchers involved are experienced and knowledgeable regarding the strategies used and the data obtained. Novice researchers will struggle to interpret divergent results. Experienced qualitative researchers, regardless of their philosophical bent, generally agree that inconsistencies are common in studies using triangulation. Rather than unraveling the credibility of the findings, divergent results often direct the researcher to a deeper appreciation for the multidimensionality of the research question and point to new avenues of inquiry.

*Sarah L. Hastings*

***See also*** Internal Validity; Mixed Methods Design; Naturalistic Inquiry; Positivism; Qualitative Research

### Further Readings

Denzin, N. K. (1978). *The research act: A theoretical introduction to sociological methods*. New York: Praeger.

Lincoln, Y. S., & Guba, E. G. (1985). Establishing trustworthiness in naturalistic inquiry. In *Naturalistic inquiry* (pp. 289–331). Beverly Hills, CA: Sage.
Morrow, S. L. (2005). Quality and trustworthiness in qualitative research in counseling psychology. *Journal of Counseling Psychology, 52,* 250–260.
Onwuegbuzie, A. (2002). Why can't we all get along? Towards a framework for unifying research paradigms. *Education, 122,* 518–530.
Patton, M. Q. (2002). Enhancing the quality and credibility of qualitative analysis. In *Qualitative research and evaluation methods* (3rd ed.) (pp. 541–588). Thousand Oaks, CA: Sage.
Shin, Y., Pender, N. J., & Yun, S. (2003). Using methodological triangulation for cultural verification of commitment to a plan for exercise scale among Korean adults with chronic diseases. *Research in Nursing and Health, 26,* 312–321.
Tashakkori, A., & Teddlie, C. (1998). *Mixed methodology: Combining qualitative and quantitative approaches.* Thousand Oaks, CA: Sage.

## Trimmed Mean

There are at least three fundamental concerns when using the mean to summarize data and compare groups. The first is that hypothesis testing methods based on the mean are known to have relatively poor power under general conditions. One of the earliest indications of why power can be poor stems from a seminal paper by John Wilder Tukey published in 1960. Roughly, the sample variance can be greatly inflated by even one unusually large or small value (called an outlier), which in turn can result in low power when using means versus other measures of central tendency that might be used. A second concern is that control over the probability of a Type I error (when the true hypothesis is wrongly rejected) can be poor. It was once thought that Type I error probabilities could be controlled reasonably well under nonnormality, but it is now known why practical problems were missed and that serious concerns can arise, even with large sample sizes. A third concern is that when dealing with skewed distributions, the mean can poorly reflect the typical response.

There are two general strategies for dealing with the problem of poor power, with advantages and disadvantages associated with both. The first is to check for outliers, discard any that are found, and use the remaining data to compute a measure of location. The second strategy is to discard a fixed proportion of the largest and smallest observations and average the rest. This latter strategy is called a trimmed mean, which includes the median as a special case, and unlike the first strategy, empirical checks on whether there are outliers are not made.

A 10% trimmed mean is computed by removing the smallest 10% and the largest 10% of the observations and averaging the values that remain. To compute a 20% trimmed mean, remove the smallest 20%, the largest 20%, and compute the usual sample mean using the data that remain. The sample mean and the usual median are special cases that represent two extremes: no trimming and the maximum amount.

A basic issue is deciding how much to trim. In terms of achieving a low standard error relative to other measures of central tendency that might be used, 20% trimming is a reasonably good choice for general use, except when sample sizes are small, in which case it has been suggested that 25% trimming be used instead. This recommendation stems in part from the goal of having a relatively small standard error when sampling from a normal distribution. The median, for example, has a relatively high standard error under normality, which in turn can mean relatively low power when testing hypotheses. This recommendation is counterintuitive based on standard training, but a simple explanation can be found in the books listed under Further Readings.

Moreover, theory and simulations indicate that as the amount of trimming increases, certain known problems with means, when testing hypotheses, diminish. This is not to say that 20% is always optimal; it is not, but it is a reasonable choice for general use. However, if too much trimming is used, with the extreme case being the median, then power can be relatively poor under normality. And when comparing medians, special methods are required for dealing with tied values.

After observations are trimmed, special methods for testing hypotheses must be used (and the same is true when discarding outliers). Moreover, appropriate estimates of standard errors must be performed in a manner that takes into account how extreme values are identified and removed.

That is, applying standard methods for means to the data that remain after trimming violates the basic principles used to derive standard techniques. Some of these methods are relatively simple to apply, whereas others are based on bootstrap methods that take advantage of modern computing power. It is now known that trimmed means also help deal with problems associated with means when the goal is to control Type I errors or compute accurate confidence intervals. A rough explanation of why trimming helps is that the sampling distribution of the trimmed mean, or some appropriate test statistic, is less affected by nonnormality versus no trimming (the sample mean). This strategy has practical importance because recent studies have found general conditions in which classic methods based on means perform poorly. For example, there are realistic conditions where the one-sample Student's $t$ test requires a sample size of 300 or more to get accurate results. The two-sample Student's $t$ test can perform poorly if the distributions differ in terms of skewness. Indeed, there are general conditions where the actual Type I error probability does not converge to the nominal level as the sample size gets large. And when comparing more than two groups, practical problems are exacerbated.

*Rand R. Wilcox*

***See also*** Descriptive Statistics; Estimation; Influential Data Points; Mean; Null Hypothesis; $p$ Value; Sampling Distributions; Significance Level, Concept of; Standard Error of Estimate; $t$ Test, One Sample; Type I Error; Type II Error; Winsorize

**Further Readings**

Rosenberger, J. L., & Gasko, M. (1983). Comparing location estimators: Trimmed means, medians, and trimean. In D. C. Hoaglin, F. Mosteller, and J. W. Tukey (Eds.), *Understanding robust and exploratory data analysis*. New York: Wiley.

Tukey, J. W. (1960). A survey of sampling from contaminated normal distributions. In I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, & H. B. Mann (Eds.), *Contributions to probability and statistics*. Stanford, CA: Stanford University Press.

Wilcox, R. R. (2001). *Fundamentals of modern statistical methods*. New York: Springer.

Wilcox, R. R. (2003). *Applying contemporary statistical techniques*. San Diego, CA: Academic Press.

Wilcox, R. R. (2009). Basic statistics: Understanding conventional methods and modern insights. Manuscript submitted for publication.

# Triple-Blind Study

Triple-blind (i.e., triple-masking) studies are randomized experiments in which the treatment or intervention is unknown to (a) the research participant, (b) the individual(s) who administer the treatment or intervention, and (c) the individual(s) who assess the outcomes. The terms *blind* and *masking* are synonymous; both terms describe methods that heslp to ensure that individuals do not know which treatment or intervention is being administered. The purpose of triple-blinding procedures is to reduce assessment bias and to increase the accuracy and objectivity of clinical outcomes.

## Examples

Conducting a triple-blind study is difficult. The following two examples highlight some challenges in conducting and evaluating triple-blind studies.

### *Example 1*

The first example, from a study by Abraham Heller and colleagues, highlights the difficulties in conducting a triple-blind study in a clinical setting of psychiatric patients ($n = 15$). The authors report using a triple-blind design to compare the effectiveness of three medications, which include imipramine, desipramine, and placebo. Participants were accepted into the study if the examiner determined that the individual was severely depressed, after which the examiner rated the individual on the Hamilton Depression Scale, and the participant completed self-reports on the Zung Self-Rating of Depression and the Minnesota Multiphasic Personality Inventory (MMPI). These scales were repeated in 1 week, biweekly during the hospitalization, and monthly for 3 months on an outpatient basis. The authors reported that a random triple-blind method was used to assign

medication to patients. Reportedly, a secretary was given the responsibility of pairing each participant's name to a list of three randomly assigned letters. After randomization, participants were monitored first in an inpatient treatment psychiatric setting, followed by treatment in an outpatient day-care setting for 1–4 weeks, and finally treatment in a regular outpatient setting. After a 3-month follow-up period, the patient was administered another MMPI and the previous scales. The authors stated that to eliminate bias no one with direct contact with the patients knew which medication the patient took during the study.

#### Challenges

This study enrolled a small sample, and the methods used do not provide sufficient detail to determine whether the clinical findings truly support use of one medication over another. For example, a secretary paired each patient's name with a letter, but it is not clear who actually administered the medications, or if the person giving the medication was truly blinded to which medication was administered. Furthermore, it is unclear how the list of letters was randomly assigned nor what methods were used to ensure that each patient did not know which medication they received. Because the side-effect profile of the three medications discussed earlier is substantially different (particularly for the placebo compared with the antidepressants), it is unlikely that the patients did not know which medication they received. It is also unclear what methods were used to ensure that the individuals that conducted the outcomes assessment did not know the clinical outcomes. There is no clear description of how the individuals who performed the outcomes were blinded to the clinical outcomes.

### *Example 2*

A second example, from a study by Seung-Jung Park and colleagues, highlights the challenges in conducting a triple-blind study using multiple clinical sites. The study seems to be a much more rigorously conducted triple-blind study compared with the Heller study discussed earlier, given the methods described to ensure a triple-blind. The study examined in symptomatic patients ($n = 177$) whether a coronary stent coated with an antiproliferative agent paclitaxel prevented restenosis (e.g., whether this agent prevented hyperplasia and reoccurrences of coronary artery occlusion). The study involved three medical centers and compared stents coated with one of two doses of paclitaxel compared to a stent that was uncoated (i.e., placebo). Randomization was conducted using blocks of participants (i.e., each block used a 1:1:1 ratio to assign participants to 1 of the 3 interventions), and each block of participants was stratified according to stent diameter. Reportedly, stents were ordered in a randomized sequence, angiography was used to determine vessel diameter, and the appropriate diameter stent was selected. The authors state that patients, investigators, and core-analysis laboratories were all unaware to what groups the patients were assigned.

#### Challenges

Although the authors used block randomization, it is unclear what procedures were used to stratify based on diameter. Also, it is unclear what methods were used to ensure the correct dosing of the paclitaxel. The authors are vague regarding who was in the core-analysis team and to what extent the core-analysis team was truly independent from the independent clinical-events committee and the data safety monitoring committee. Furthermore, it is not known what procedures were in place to ensure blinding of results across the three clinical sites.

## Conclusion

Although the triple-blind study offers clear advantages of control from various sources of bias, these designs are logistically difficult to conduct. Prior to initiating a triple-blind study, a conscientious effort needs to be made to identify the necessary parties who will conduct the study. There should be clear guidelines in place regarding which individuals and monitoring boards will be blinded and how investigators will maintain the triple blind. In evaluating any study that reports using a triple blind, a critical eye is needed to determine whether triple-blind procedures were rigorously enforced.

*Michael A. Dawes*

*See also* Randomization Tests; Research Design Principles; Research Hypothesis; Threats to Validity

**Further Readings**

Heller, A., Zahourek, R., & Whittington, H. G. (1971). Effectiveness of antidepressant drugs: A triple-blind study comparing imipramine, desipramine, and placebo. *American Journal of Psychiatry, 127,* 132–135.

Park, S. J., Shim, W. H., Ho, D. S., Raizner, A. E., Park, S. W., Hong, M. K., et al. (2003). A paclitaxel-eluting stent for the prevention of coronary restenosis. *New England Journal of Medicine, 348,* 1537–1545.

Piantodosi, S. (2005). *Clinical trials: A methodological perspective* (pp. 180–181). Hoboken, NJ: Wiley.

Schuster, D. P., & Powers, W. J. (2005). *Translational and experimental clinical research* (p. 95). Philadelphia: Lippincott Williams & Wilkins.

# True Experimental Design

The exact definition of true experimental designs has been debated. The term *true experiment* is sometimes used to refer to any randomized experiment. In other instances, the term *true experiment* is used to describe all studies with at least one independent variable that is experimentally manipulated and with at least one dependent or outcome variable. Furthermore, the word *true* has been interpreted to mean there are a limited number of correct experimental methods. Regardless of the exact definition, the distinguishing feature of true experimental designs is that the units of study—be they individuals, animals, time periods, areas, clinics, or institutions (i.e., whoever or whatever is assigned to an experimental condition is the unit)—are randomly assigned to different treatment conditions.

Because of the various meanings and the use of the term *true experiment,* others have preferred to use the term *randomized experiment* as a clearer, more concise description of the experimental design. A randomized experiment is any experiment in which units are randomly assigned to receive one or more treatments or one or more alternative control conditions. Randomized assignment refers to any procedure for assigning units to either treatment or control conditions based on chance, with the potential for every unit to be assigned to one of the treatment conditions at a greater than zero probability. In the following sections, how random assignment facilitates causal inference is discussed, and common designs that use random assignment are presented.

## Random Assignment and Causality

For randomized experiments, the use of random assignment is important because it facilitates causal inference. For example, randomization ensures that the experimental units' treatment condition is not confounded by an alternative cause or systematically introduced difference between conditions. Randomization, therefore, reduces potential threats to experimental validity by dispersing these threats randomly across experimental groups to minimize group differences before treatment begins. Randomization allows the researcher to know how groups were assigned and has the advantage that the selection process can be modeled accurately. Random assignment also allows for the computation of valid estimates of error variance that are orthogonal to treatment conditions. Each advantage captures different aspects of the benefits of randomization. Random assignment is the only design feature that accomplishes all the above in the same experiment, and it does so reliably while minimizing threats to internal validity.

## Designs Using Random Assignment

This section presents various randomized experimental designs. The following designs are among the most commonly used in field and treatment research and are basic designs from which more complex designs can be built. With each design, key features include control and manipulation over the timing, intensity, and duration of experimental variables. The basic design of a randomized experiment requires a minimum of two conditions, with random assignment of experimental units to treatment conditions, followed by posttest assessment of units. Herein, the letter "R" indicates that the group on a given line is formed by random assignment. Although random assignment R is placed at the beginning of each line, in practice,

randomization often occurs before or after pretests. "X" is the treatment and "O" is the observation. The basic design can be depicted as

R X O
R O

The diagram presented above depicts a two-group experiment in which one group receives the experimental treatment X followed by observation O, and the second group acts as a comparison to be observed without experimental manipulation. This second group is one example of a control condition. An important issue with this design is the selection of the control condition. In all randomized experiments, the researcher must consider what factors need to be controlled in the design of the study. Depending on what hypothesis needs to be tested, the number and type of treatment and control groups might vary. Examples of control groups include no-treatment controls, dose-response controls, wait-list controls, expectancy controls, or attention-only controls.

### *The Basic Randomized Design Comparing Two Treatments*

One variation of the randomized design uses two treatments, one of which is a gold standard treatment of known efficacy. This design can be used to test what is the effect of the new treatment compared with the effect of the standard treatment. This design often works well when the standard treatment has already been shown to be effective when compared with no-treatment controls (see basic design, presented previously):

R $X_A$ O
R $X_B$ O

### *The Basic Randomized Design Comparing Two Treatments and a Control*

A second variation of the randomized design is used when it is unknown whether experimental units after treatments A and B are likely to be different (e.g., because of the treatment being equally effective or equally ineffective) at posttest. In this case, a control group is added to the basic randomized design, as follows:

R $X_A$ O
R $X_B$ O
R O

### *The Pretest–Posttest Control Group Design*

Often, it is advantageous to add a pretest to the basic randomized design, as follows:

R O X O
R O O

Or, random assignment can occur after the pretest, as follows:

O R X O
O R O

The two designs presented earlier are frequently used for randomized field experiments because they offer two major advantages. First, they can help to minimize threats to internal validity resulting from attrition. Second, they can be used with specific statistics to increase power to reject the null hypothesis.

In some instances, however, it makes sense not to use a pretest. Examples include when pretesting is expected to produce an unwanted change in the experimental units before the treatment occurs, when pretesting cannot be obtained, and when the pretest condition is known to be constant.

### *The Alternative-Treatment Design with Pretest*

Pretests are also recommended when comparing different treatments. If the researcher has reason to believe there might be no difference in outcomes between two alternative treatments, this design allows for the determination of whether both groups improved or whether neither did. Another advantage of this design is its utility when the researcher faces the challenge that it is unethical to withhold treatment to a group of patients, but it is unknown which treatment is more effective. Alternatively, this design can be used when there is a gold-standard treatment against which all other treatments can be compared.

R O $X_A$ O
R O $X_B$ O

### *Multiple Treatments and Controls With Pretest*

Randomized experiments with pretests can also include a control group and multiple treatment groups. This design can be expanded to include more than two treatments and more than one control condition. For example, the Treatment of Depression Collaborative Research Program studied 250 depressed adults who were randomized to receive one of three active treatments: cognitive behavior therapy, interpersonal therapy, antidepressant therapy (imipramine) with clinical management, or placebo medication with clinical management (i.e., control condition).

R O $X_A$ O
R O $X_B$ O
R O $X_C$ O
R O O

This design can also be used to manipulate experimentally independent variables that are likely to have differing measurable effects on experimental variables at different doses or with increasing levels (i.e., parametric or dose-response studies). For example, in a study conducted in Cali, Columbia, adults were administered educational, nutritional, and medical treatment in four "doses" of total time receiving the following treatments: 990 hours, 2,070 hours, 3,130 hours, and 4,170 hours. The results showed significant effects at higher doses, with minimal effects at the lowest dosage.

### *Factorial Designs*

Factorial designs use two or more independent variables (i.e., factors), which have at least two levels, and are combined such that all logical combinations of factors can be tested. The simplest version of this design is called a two-by-two factorial design, which is depicted as follows:

R $X_{A_1B_1}$ O
R $X_{A_1B_2}$ O
R $X_{A_2B_1}$ O
R $X_{A_2B_2}$ O

Factorial designs have several advantages. They can allow for smaller sample sizes than might be needed when using other designs, because each participant can receive the treatment conditions simultaneously. Second, factorial designs can be used to test whether a combination of treatments is more effective than one treatment or no treatment. Third, experiments that use a factorial design can be used to test for interactions (i.e., when the effects of treatment vary over levels of another variable).

Although factorial designs have clear advantages and are commonly used in laboratory research, they also have disadvantages because of the difficulty of implementing this design in many situations. For example, factorial designs require rigorous control over the combinations of treatments administered to each experimental unit, but maintaining control over multiple factors and multiple levels can pose special challenges.

### *Longitudinal Designs*

Longitudinal designs use multiple observations that occur before, during, and after treatment, with the number and timing of observations determined by the specific hypotheses. Longitudinal designs have several advantages. They can determine change over time with the use of growth-curve models of individual differences in response to treatment. Longitudinal designs can also increase statistical power when five or more waves of observations are used, even when sample sizes are small.

R O...O X O O...O
R O...O O O...O

Unfortunately, longitudinal designs have many disadvantages. Importantly, attrition might increase with longer follow-up periods and require rigorous follow-up procedures. For longitudinal randomized treatment studies, it is often unethical to withhold treatment for long periods of time using no-treatment or wait-list control groups.

### *Crossover Designs*

Crossover designs are generally used with participants who are randomized to receive two or more treatments, after which they receive a posttest. After the posttest, the participants are assigned to the other treatment they did not receive previously and then they receive a second posttest, in the following design:

R O $X_A$ O $X_B$ O

R O $X_B$ O $X_A$ O

The crossover design is used to counterbalance and assess order effects. The crossover design is also used to obtain more information on causality than would be possible with only the first posttests. In either case, crossover designs are most useful when the treatments do not produce carryover effects. (Carryover effects occur when one treatment has residual effects such that the effects of the first treatment continue after beginning a second treatment.)

### *Solomon Four-Group Design*

The Solomon four-group design is often used to control for effects of pretesting. This design uses two experimental groups and two control groups. After randomization, for each respective group, one receives a pretest and the other does not. This design allows for examination of the effects of the pretest to determine whether the pretest itself is an active treatment or intervention.

R O $X_1$ O

R – $X_1$ O

R O $X_2$ O

R – $X_2$ O

## Conclusion

The *true experimental design* is a synonym for randomized experiments. Randomized experiments are useful because they facilitate causal inference and reduce threats to internal validity. A variety of experimental designs can be used, depending on what research hypotheses are to be tested. However, randomized experiments are often complex, expensive, and difficult to do. Therefore, they should be conducted only after preliminary study results support very specific testable hypotheses.

*Michael A. Dawes*

***See also*** Cause and Effect; Internal Validity; Quasi-Experimental Design; Randomization Tests; Research Design Principles; Research Hypothesis; Threats to Validity

### Further Readings

Bauman, K. E. (1997). The effectiveness of family planning programs evaluated with true experimental designs. *American Journal of Public Health, 87,* 666–669.

Elkin, L., Shea, T., Watkins, J. T., Imber, S. D., Sotsky, S. M., Collins, J. F., et al. (1989). National Institute of Mental Health Treatment of Depression collaborative research program: General effectiveness of treatments. *Archives of General Psychiatry, 46,* 971–982.

Imber, S. D., Pilkonis, P. A., Sotsky, S. M., Elkin, I., Watkins, J. T., Collins, J. F., et al. (1990). Mode-specific effects among three treatments of depression. *Journal of Consulting and Clinical Psychology, 58,* 352–359.

Maxwell, S. E., & Delaney, H.D. (1990). *Designing experiments and analyzing data: A model comparison approach*. Pacific Grove, CA: Brooks/Cole.

McKay, H., Sinisterra, L., McKay, A., Gomez, H., & Lloreda, P. (1978). Improving cognitive ability in chronically deprived children. *Science, 200,* 270–278.

Meltzoff, J. (2006). *Critical thinking about research: Psychology and related fields* (p. 89). Washington, DC: American Psychological Association.

Ribisl, K. M., Walton, M. A., Mowbray, C. T., Luke, D. A., Davidson, W. S., & Bootsmiller, B. J. (1996). Minimizing participant attrition in panel studies through the use of effective retention and tracking strategies: Review and recommendations. *Evaluation and Program Planning, 19,* 1–25.

Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis*. New York: McGraw-Hill.

Shadish, W. R., Cook, T. D, & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin.

# True Positive

True positive, as it pertains to null hypothesis significance testing (also known as hypothesis

**Table 1** Statistical Decision Table: The Four Potential Results with Any Null Hypothesis Significance Test

| | | *In Reality* | |
|---|---|---|---|
| | | **Null hypothesis is true**<br>*There are no differences between the groups* | **Null hypothesis is false**<br>*There are differences between the groups* |
| *Results of the Research Study* | **Fail to reject the null hypothesis**<br>*Find no differences between the groups* | Correct decision<br>**True negative** | Type I error<br>**False negative** |
| | **Reject the null hypothesis**<br>*Find differences between the groups* | Type II error<br>**False positive** | Correct decision<br>**True positive** |

testing), is correctly rejecting the null hypothesis when it is in fact false. In other words, it is correctly finding a correlation or a difference between groups. Research design is founded on the principles of hypothesis testing and in short, the pursuit of discovering true positives is a goal. In this entry, the definition of the true positive is developed in the context of null hypothesis significance testing, how it relates to power analysis is described, and an example is provided.

## Definition

Null hypothesis significance testing examines the probability of two outcomes (the null and alternative hypothesis). Because these two hypotheses are mutually exclusive and exhaustive, there are four potential results. Two of these results are correct decisions. Two potential results are incorrect decisions. A true positive is one of the potential correct decisions. It is finding a difference between groups (e.g., males and females) or a relationship (e.g., the association between height and weight) that truly exists and does not represent a chance fluctuation. The table below outlines the four possible conditions and states the research question in terms of group differences. The same table can be made for predictive or correlational relationships.

It is desirable to find a difference between the groups (or an association between the variables) when it truly exists. In other words, the goal of a research project is to detect the true state of affairs. If men and women are different with respect to height, then the researcher wants to be able to reproduce this difference. The concept of a true positive (and the concepts of Type I error, Type II error, and true negative) is a theoretical abstraction. The true state of affairs can never really be known (because of random error, measurement error, etc.). Thus, it is not possible to determine whether an error has been made or whether a true positive or true negative exists. The use of statistical significance testing attempts to limit the likelihood of making an error and to increase the chance that the correct decision is made.

The word *positive* in this case does not reference *good*. The positive with respect to the true positive is more akin to the presence (versus the absence) of a condition or relationship (e.g., having chicken pox). Therefore, a true positive indicates that condition, association, or difference exists.

## In Relation to Power Analysis

Another concept that is related to the idea of true positive is the concept of statistical power. Statistical power analysis is the sensitivity of a research design to detect a true positive. In other words, it is the ability of the research study to find a difference or reject the null hypothesis when it is in fact false. Statistical power and alpha level (the likelihood of a false negative or a Type I error) are impacted by the design of the research study and influence each other. Statistical power is determined by the effect size, directional nature of the test, and the alpha level. The alpha level is set a priori (prior to running the research study) by the researcher.

## Example

To illustrate the concept of the true positive, the case of Naomi's potential pregnancy is provided. Naomi wants to determine whether she is

pregnant. She takes a pregnancy test (pregnancy was chosen because the reality of the situation will eventually be known in contrast to real research situations). After following the directions correctly, Naomi gets a positive pregnancy test (indicating that she is pregnant). As the weeks follow, Naomi begins to show more signs of the pregnancy (i.e., weight gain, hearing the baby's heart beat, and feeling the baby move) and eventually gives birth. In short, Naomi's pregnancy test result was a true positive. The test said she was pregnant and she really was.

In contrast, Naomi's pregnancy test could have said that she was pregnant and she really was not (i.e., false positive or false report). In this case, she could start preparing for the pregnancy only to later find out that the test was wrong. The test could have also said that she was not pregnant. In that case, if she had really been pregnant, it would have been a false negative. As with any error, there are repercussions for the false information. Naomi could have thought she was not pregnant and decided to drink alcohol or participate in other activities that are considered dangerous for pregnant women. If the test said that Naomi was not pregnant and in reality, she was not, then it was a correct result.

*Rose Marie Ward*

***See also*** Alternative Hypotheses; Coefficient Alpha; Null Hypothesis; Power Analysis; Type I Error; Type II Error

**Further Readings**

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997–1003.

Mulaik, S. A., Raju, N. S., & Harshman, R. A. (1997). There is a time and place for significance testing. In L. A. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 65–116). Mahwah, NJ: Lawrence Erlbaum.

Neyman, J., & Pearson, E. S. (1967). *Joint statistical papers.* Cambridge, MA: Cambridge University Press.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5,* 241–301.

Scarr, S. (1997). Rules of evidence: A larger context for the statistical debate. *Psychological Science, 8,* 16–17.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.

# True Score

True score, which is the primary element of true score theory, is the individual's score on a measure if there was no error. Some classic theories of measurement believe that a true score can be estimated through repeated testing. The concept of true score is important to research design as it emphasizes that there is some error involved in any type of measurement (e.g., height, weight, self-esteem, IQ, and heart rate). In this entry, the definition of a true score is developed and explored with respect to reliability, measurement error, and classic extensions. Finally, some alternatives to true score theory are briefly presented.

## Definition

True score can also be defined with respect to the idea of an observed score. In short, the observed score is the true score plus some error. Symbolically, the equation is

$$X_o = X_t + E,$$

where $X_o$ is the observed score or the participant's score on a measure (e.g., the participant's score on Prochaska's smoking decisional balance inventory) and $X_t$ is the participants true score or what is really of interest. In addition, E represents the error term that can be comprised of measurement error (i.e., systematic error) and some random error (i.e., unsystematic error). For example, real body weight (i.e., the true score of our weight) is never really known as it fluctuates with time of day, gravitational pull, retention of water, clothing choice, and many other variables. The number viewed on a scale would be considered an observed score (the true weight plus some error). The observed weight on the scale is based on the real weight with the addition of the calibration of the instrument used to measure weight plus the other variables that might interfere with the scales ability to properly measure weight. Conceptually, the formula would be

$$\text{Observed score} = \text{true score} + (\text{measurement error} + \text{random error})$$

or

$$\text{Weight on the scale} = \text{real weight} + \text{calibration of scale} + \text{time of day, etc.}$$

Although weight can be measured on a scale, it is important to remember that it is estimated. Without knowing the amount of error or the true score, the observed weight is actually meaningless. The obtained weight (or measurement on a myriad of other psychological measures or tests) could be greatly flawed (i.e., have a large amount of error) or very precise (i.e., have a little amount of error). Without multiple measurements on the individual or test subject or replication studies, it is difficult to know how much error is in the measurement.

## Assumptions of True Score Theory

Because of the theoretical nature of the true score, $X_t$, it is not possible to know its value. Although it cannot be directly measured, there are some basic assumptions (according to true score theory and classic measurement theory) concerning the true score. As mentioned earlier, the first assumption is that all observed variables contain some error. If there was no error, then the true score would equal the observed score $X_o$. Another assumption is that the true score and the error score are independent. In other words, the correlation between the true score and the error score is zero (signifying that the movement in the error score has no impact on the value of the true score). The error score $E$ is considered completely random. For example, if a baby weighs 12 pounds (her true score) and is weighed multiple times, then the amount the baby's weight fluctuates (thereby showing the differential error values) does not impact the baby's real weight.

Beyond the error of measurement and the lack of relationship between the true score and the error score, it is also assumed the mean of the error scores will be equal to zero. Some error scores will be above the true score and some error scores will be below the true score. It is assumed that the pattern of the error scores (over repeated testing) will represent the normal curve and thus have a mean of zero. Another assumption is that the expected value of $X_o$ is $X_t$. In other words, the observed variable is to be a good estimate of the true score. This assumption could be difficult in the case of someone who is unfamiliar with the language used in a measure. For example, if the participant's primary language is not English and the measure is constructed in English, then he or she might miss the true meaning of the items and therefore provide answers that do not reflect his or her true ability or disposition. In true score theory, it is also assumed that the error terms across tests will not be correlated. For example, if participants are completing a battery of measures (e.g., emotional intelligence, alcohol use, or self-efficacy), the error terms on each of these measures will not be related. This assumption is difficult to obtain if the participant's mood (or fatigue or other variable) impacts his or her ability to complete the measure. True score theory presupposes that the error on one measure or test will not be related to the true score on another test (i.e., the correlation between $E_1$ and $X_{t2}$ is zero). These assumptions provide the basis for reliability testing.

## Reliability Testing

As mentioned earlier, replication is necessary to estimate a true score. There are two types of replication. The first type of replication is test–retest. It involves assessing the participants or research subjects multiple times using the same measure (stability over time), for example, giving participants a measure of emotional intelligence today and giving them the same measure of emotional intelligence next week. Because it is assumed that the true score is stable (or trait-like), the scores of the participants could be correlated. The correlation would indicate the stability or reliability of the measure over time. Another type of replication is measuring emotional intelligence across different participants.

Several tests for the reliability of measures were developed based on the concepts of true score theory. G. Frederic Kuder and Marion Richardson's KR-20 and KR-21 were used to provide statistical estimates of reliability. The most popular method of estimating reliability is Lee Cronbach's

coefficient alpha (which is quantified on a scale of 0 to 1, with higher scores indicating more internal consistency). It is a measure of the internal consistency of a series of items using the observed scores of the participants. In short, it is the variance of the observed variable that is ascribed to the stable true scores. The variation in the observed scores caused by error (as error decreases reliability increases) is discussed in the next section.

## Error

If it were possible to know the true score, then it could be subtracted from the observed score to obtain the error score. In essence, the error score represents the observed score's deviation from the true score. As mentioned earlier, it can be composed of two components (systematic and unsystematic error). Although some classic theories only address unsystematic error (also known as random error), this section develops both concepts with respect to true score theory.

The two types of error affect the observed variable in different ways. Systematic error effects all of the measurements equally (i.e., a constant is added to all scores). An example of a systematic error is in the case of two professors team teaching the same class in the same manner. If professor A consistently grades the students 10 points higher than professor B, then there is a systematic error in the grades of the students. Another example of a systematic error would be if professor A graded the left-handed students easier than the right-handed students (in other words, there was a bias against right-handed students). Unsystematic (or random) errors complicate the relationship between the true score and the observed score.

Another concept important to error and true scores is the standard error of measurement. Similar to the concept of standard deviation when the word *standard* indicates a mathematical average, standard error of measurement is an average of the error terms about the true score. In short, it is an estimate of how different the observed scores are from the true score on average. The standard error of measurement is commonly used to develop confidence intervals for the observed scores. The confidence interval surrounds the true score. In other words, a 95% confidence interval indicates that there is a 95% confidence that the participant's true score falls within the range of the confidence interval (e.g., 95% confident that the participant's IQ falls between 70 and 130 on a standard measure of intelligence).

## Extensions of True Score Theory

Measurement error is a key term in true score theory. If the standard error of measurements are equivalent across two tests, the true scores are equivalent ($X_{t1} = X_{t2}$), and the tests do not violate the aforementioned assumptions, then the two tests are considered parallel forms. Parallel forms are in many educational testing situations (e.g., GREs, SATs, ACTs, and LSATs). Another extension of true score theory is the concept of tau equivalent tests. In essence, tau equivalent tests are tests that the true scores differ by a constant ($X_{t1} = X_{t2} + c$). Tau equivalent tests are not necessarily parallel tests.

## Alternatives to True Score Theory

Beyond the extensions of the true score theory, theorists have suggested some alternative theories or models. One alternative is domain sampling. It was developed in rebuttal of the rigid dichotomy of true and error scores. The underlying idea of domain sampling is that a constellation of behaviors can represent a trait or domain. For example, exercise is a broad domain in which many people choose different behaviors (e.g., running, walking, swimming, tennis, rowing, and biking). To measure the trait or domain, a series of items or questions measures different aspects of the domain. Within the concept of domain sampling, the idea of parallel tests that was previously mentioned is an extension or a special case. Another special case of domain sampling is factorial domain-sampling. In this case, within a domain, there can be several factors (or subcategories). For example, within the previous example of exercise, factors could be weight-bearing exercises (e.g., running and walking) versus non-weight-bearing exercises (rowing and biking). Another alternative to true score theory is the binomial model. In this model, it is assumed that the error scores have a binomial distribution around the true scores. Contemporary alternatives include latent variable modeling (supposing that there is an underlying trait to a series

of items) and multitrait multimethod (using multiple venues to access an estimation of a trait). A discussion of these methods is beyond the current entry. A final alternative is generalizability theory. It uses a universe score (similar to the true score in that it is an expected value for the observed score across all valid observations) and clearly outlines the dimensions in which a test can be used.

*Rose Marie Ward*

***See also*** Coefficient Alpha; Classical Test Theory; Error; Reliability

**Further Readings**

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.
Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.
Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.
Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology, 3,* 1–18.
Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.
Rao, C. R., & Sinharay, S. (2007). *Handbook of statistics 26.* Boston: Elsevier.
Thurstone, L. L. (1931). Measurement of social attitudes. *Journal of Abnormal and Social Psychology, 26,* 249–269.
Tyron, R. C. (1957). Reliability and behavior domain validity: Reformulation and historical critique. *Psychological Bulletin, 54,* 229–249.

# *t* Test, Independent Samples

An independent samples *t* test is a hypothesis test for determining whether the population means of two independent groups are the same. The researcher begins by selecting a sample of observations and estimates the population mean of each group from the sample means. The researcher then compares these two sample means via the formula

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}},$$

where $\overline{x}_1$ is the sample mean of group 1, $\overline{x}_2$ is the sample mean of group 2, $s_p^2$ is the pooled estimate of variance (formula given in the example problem), $n_1$ is the sample size of group 1, and $n_2$ is the sample size of group 2. This *t* value can then be used to determine the likelihood that any difference between the two sample means is real versus caused by chance.

For example, assume a researcher is interested in the effect of stress on students' performance. She randomly assigns participants to either a "stress" or "no stress" group and examines how each of the two groups performs on a standardized test by calculating the sample mean for each group. Because the researcher is not interested in the samples per se but in all students, she needs to determine whether any difference between her sample means is real or a result of chance. The independent samples *t* test will assist her in making this determination.

Note that the equation presented earlier tests whether the population means are different. It is also possible to use the independent samples *t* test to examine whether the difference between the population means is greater than a constant, c (rather than 0). In this situation, the numerator of the formula for *t* changes to $(\overline{x}_1 - \overline{x}_2 - c)$. In most cases, however, the independent samples *t* test is used to determine whether the population means are equal, and the rest of the discussion is confined to that situation.

## Uses

The independent samples *t* test is appropriate whenever the researcher wants to know whether two population group means are different and when the observations in each of the groups are independent of the observations in the other group. In some cases, these groups are composed from naturally occurring populations. For example, a researcher might be interested in comparing the spatial ability of men and women. In this case, all men and women in the country (or at his or her university, etc.) would comprise the two populations (one of men and one of women). The researcher would then construct a sample of men and of women, calculate the sample means, and conduct an independent samples *t* test to

determine whether any differences between the sample means are real or caused by chance.

In many cases, however, the populations are not naturally occurring but instead are determined by some element of the study. This is particularly common in experimental research, where the researcher manipulates a variable (stress in the preceding example) and assigns participants to one of the two resulting groups. In this case, the groups that result from the manipulation do not reflect actual known populations. Instead, they reflect hypothetical infinite populations that result from the experimental manipulation.

The conclusions that can be drawn from the independent samples *t* test are somewhat different depending on whether the populations are naturally occurring or result from an experimental manipulation (for example, it is easier to demonstrate causality when a variable has been manipulated). Nonetheless, the independent samples *t* test functions identically in the two situations, testing whether the population means of two independent groups are equal.

## Interpreting the *t* Value

The result of an independent samples *t* test is a *t* value. This *t* value is compared with the *t*-distribution that would occur if the two population group means were equal. This possibility (that the two population group means are equal) is known as the null hypothesis. The logic that underlies the independent samples *t* test is to compare the *observed* value of *t* (that is, the value of *t* calculated from the independent samples *t* test) with the *t*-distribution that would occur if the null hypothesis were true. If the observed value of *t* is in keeping with this distribution, then the null hypothesis might well be true (i.e., the population means are equal). If the observed value of *t* is not in keeping with this distribution, then the null hypothesis is likely *not* true (i.e., the population means are not equal).

The precise form of the *t*-distribution is determined by degrees of freedom (*df*), which in turn is determined by the sample size (specifically, $df = n_1 + n_2 - 2$). Basically, the larger the number of observations, the better one is able to estimate the variability in the data, so the *t*-distribution most closely resembles the *z* distribution with a large number of observations.

The *t*-distribution for 100 *df* is presented in Figure 1. Intuitively, the observed value of *t* reflects how different from each other the sample means are. Since the absolute distance between the means can be determined by many factors (e.g., the range of possible scores, size of one's sample), this difference is "standardized" by $\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$, which is the standard error of the difference between the means. Thus, the value of *t* indicates how far apart the sample means are, in standardized form. The *t*-distribution (such as the one presented in Figure 1) provides the range of possible *t* values that would reasonably be expected to be found, *assuming the null hypothesis is true* (i.e., assuming the population means are equal). Put differently, even if the population means are equal, the *sample* means will likely be different just by chance. The *t*-distribution describes precisely how different the sample means would be just by chance.

To determine whether the population means could be equal, then, one compares the *observed t* value with the *t*-distribution constructed under the assumption that the population means are equal. Assume, for example, that the observed *t* value is .50. As can be observed in Figure 1, there is a high probability of obtaining this *t* value if the population means are equal. If, however, the observed *t* value were 4.0, then there is a near-zero probability of obtaining this *t* value if the population means are equal. Thus, in that latter situation, the researcher would conclude that the population means are *not* equal.

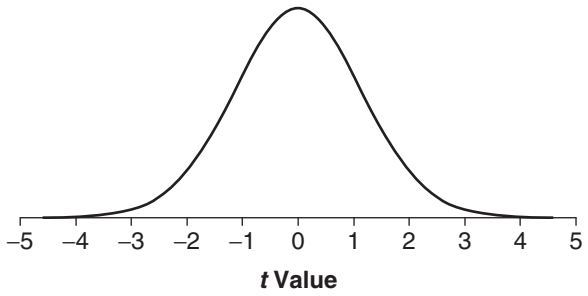



**Figure 1** *t*-Distribution with 100 *df*

In the previous examples, it was relatively straightforward to determine visually whether the observed value of *t* was in keeping with the *t*-distribution. A more precise indication is provided by a *p* value, which provides the likelihood of obtaining the observed *t* value by chance (i.e., assuming the null hypothesis is true). If the *p* value is low, then the chance of having obtained one's results just by chance (i.e., if the population means really are equal) is low, and the researcher concludes that the population means are likely not equal. In this situation, the result of the *t* test is often described as "statistically significant," because it indicates that any difference between the sample means is real and not just a result of chance. Conversely, if the *p* value is high, then the likelihood of having obtained one's results just by chance is high, and the researcher concludes that the population means might be equal. In many disciplines, .05 is used as a cutoff to determine whether the *p* value is low or high. Using this cutoff of .05 ensures that—of all times when the population means are equal—the researcher will only erroneously conclude that the population means are *not* equal 5% of the time. In statistical terms, the researcher will only make a Type I error 5% of the time when the null hypothesis is true.

The exact *p* value is provided by any standard statistical program as a result of conducting an independent samples *t* test. If, however, the independent samples *t* test is conducted by hand, calculation of the *p* value is not straightforward. Thus, the observed *t* value is sometimes compared with a "critical" value of *t*. This critical *t* value is the value of *t* that would produce a *p* value of .05. Thus, *t* values greater than the critical value of *t* produce *p* values less than .05 and lead to the rejection of the null hypothesis (concluding that the population means are not equal), and *t* values less than the critical value of *t* produce *p* values greater than .05 and lead to acceptance of the null hypothesis (concluding that the population means are equal). Critical values of *t* can be found in many statistical textbooks.

## One-Tailed Versus Two-Tailed Tests

The exact method for calculating the *p* value (and critical value of *t*) depends on whether one is conducting a directional (one-tailed) test or a nondirectional (two-tailed) test. For a nondirectional test, the researcher tests whether the population means are different (i.e., whether either mean is greater than the other one). In the previous stress example, for instance, the researcher might be interested in determining whether stress increases or decreases performance. In that situation, the researcher would perform a nondirectional, two-tailed test. Conversely, if the researcher were interested in testing whether stress *decreases* performance, the researcher would conduct a directional, one-tailed test.

The choice of whether to perform a one-tailed or two-tailed test influences the resulting *p* value. As long as the results come out in the predicted direction, the one-tailed test will produce a smaller *p* value (and, equivalently, a smaller critical value of *t*) than will a two-tailed test, thus making it easier to find statistically significant results. However, it is important to emphasize that the decision to conduct a one-tailed test must be made for theoretical reasons and prior to an examination of the data. If, for example, the researcher predicted that stress would decrease performance and found that the sample mean of the stress group was greater than the sample mean of the no-stress group, the researcher must conclude that his or her hypothesis was not supported, rather than conduct a two-tailed test. For this reason, it is sometimes recommended that researchers typically conduct two-tailed tests, to allow for conclusions on either side of the *t*-distribution.

## Assumptions

The three primary requirements for the independent samples *t* test to produce accurate results are as follows: independent observations, equal variances in the two groups, and normality. In addition, when the populations are naturally occurring (rather than resulting from an experimental manipulation), it is important to have unbiased sampling from the populations of interest.

The first assumption is independent observations. Observations in a study are *independent* to the extent that each observation is not predictive of another observation in the study. For example, assume that for ease of conducting the study, our researcher interested in the effect of stress on standardized test performance ran a session testing

20 participants in the stress condition, then 20 participants in the no stress condition, and so on. Also assume that in one of the stress sessions, one participant raced through the test, finished quickly, and left early. If the other participants in that session responded by going through the test more quickly than they would have otherwise (and thus did not perform as well), the assumption of independent observations would be violated. Note that the difficulty here is not with the person who raced through the test—this person's poor score would just constitute random error in the study. The problem occurs when the other 19 people's scores in the stress condition are affected as well, thus meaning that each observation is not independent of each other and presumably leading to worse performance in that condition.

The second assumption of the independent samples *t* test is homogeneity (equality) of variance in each of the populations. Part of the formula for calculating the *t* value involves calculating $s_p^2$, which is the pooled estimate of variance. This pooled estimate of variance reflects the best estimate of the population variance, which is assumed to be equal for each of the two groups. To determine whether the two groups actually have equal variances, one approach is to conduct Levene's test, which is provided in many statistical packages. Finding significance with Levene's test indicates that the two variances are *not* equal, although as with all hypothesis tests, a variety of factors (e.g., sample size) will influence the likelihood of achieving significant results. An alternative approach is to compare the ratio of the sample variances of the two groups. Although these should differ to some extent just by chance, one recommendation is that if the ratio of the variances (note: this is variances, not standard deviations) is more than 4:1 for small sample sizes or 2:1 for large sample sizes, then the assumption is likely violated. Note, however, that as long as the sample sizes in each of the two groups are roughly the same, violation of the equal variance assumption seems to have little influence on the results. Thus, it is generally recommended not to be concerned with this assumption unless the sample sizes in each of the groups are substantially different. If that is the case and the assumption seems to be violated, then there are multiple options for handling this situation. One option is to conduct some type of transformation on the data. A second option is to conduct a hypothesis test that does not require the equality of variance assumption (e.g., to use a modified *t* statistic that does not require calculation of $s_p^2$ or a nonparametric equivalent such as the Mann–Whitney *U* test).

The third assumption of the independent samples *t* test is that each sample was drawn from a population that follows a normal distribution. Partly because of the Central Limit Theorem, however, this assumption can typically be violated as long as the sample sizes are reasonably large. "Reasonably large" is often defined as $n > 30$; however, as long as the departure from normality is not too severe, considerably smaller sample sizes are acceptable.

Finally, if the populations are naturally occurring (rather than resulting from an experimental manipulation), it is important to have unbiased sampling from the populations of interest. Note this requirement must be met for the results of the study to be meaningful descriptively as well as inferentially. If unbiased sampling is not used, then the researcher might find statistically significant results not because the population means are actually different but instead because of the biased sampling procedures used.

## Power

As with any hypothesis test, even when the population group means are different from one another, the independent samples *t* test might not detect this difference (i.e., the results might not be statistically significant). The ability to find statistically significant results when there *is* a real difference between the population means is referred to as the *power* of the test. It is beneficial to conduct one's study with as high a power as possible, to provide the greatest chance of being able to detect a difference if one exists.

Given a certain alpha level and choice of one-tailed versus two-tailed test, the following three factors influence the power of the independent samples *t* test: the difference between the population means (larger differences = greater power), the amount of variability in the population (smaller variability = greater power), and the size of the samples (larger sample size = greater power). Jacob Cohen provides statistical tables that can be

used to determine one's power and to determine how large a sample size is needed to provide a given power.

## Example Problem

A pharmaceutical researcher is interested in determining which of two drugs is more effective at reducing cholesterol levels. He enlists the participation of 18 participants who have high cholesterol and randomly assigns 9 of them to receive Drug A and the other 9 to receive Drug B. At the end of one year of treatment, he measures their cholesterol levels, which are provided below:

Drug A: 250, 200, 230, 260, 180, 250, 290, 220, 190

Drug B: 260, 220, 170, 160, 250, 190, 250, 190, 200

The first step in testing his hypothesis is to calculate the sample means for each group, which are

$$\overline{x}_{\text{DrugA}} = \frac{\sum_{i=1}^{n} x_i}{n_{\text{DrugA}}} = \frac{2070}{9} = 230,$$

$$\overline{x}_{\text{DrugB}} = \frac{\sum_{i=1}^{n} x_i}{n_{\text{DrugB}}} = \frac{1890}{9} = 210.$$

Just based on the sample means, it seems that Drug B was more effective than Drug A at producing low cholesterol levels. However, this difference could have occurred just by chance. To examine this issue, the researcher would conduct an independent samples *t* test. The formula for the independent samples *t* test applied to this situation is:

$$t = \frac{\overline{x}_{\text{DrugA}} - \overline{x}_{\text{DrugB}}}{\sqrt{\frac{s_p^2}{n_{\text{DrugA}}} + \frac{s_p^2}{n_{\text{DrugB}}}}}.$$

The next step is to calculate the pooled estimate of variance. The formula for $s_p^2$ is

$$s_p^2 = \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}, \text{ or in this case,}$$

$$s_p^2 = \frac{s_{\text{DrugA}}^2(n_{\text{DrugA}} - 1) + s_{\text{DrugB}}^2(n_{\text{DrugB}} - 1)}{n_{\text{DrugA}} + n_{\text{DrugB}} - 2}.$$

The variance of each of the groups is

$$s_{\text{DrugA}}^2 = \frac{\sum_{i}^{n} (x_i - \overline{x}_{\text{DrugA}})^2}{n_{\text{DrugA}} - 1} = 1300,$$

$$s_{\text{DrugB}}^2 = \frac{\sum_{i}^{n} (x_i - \overline{x}_{\text{DrugB}})^2}{n_{\text{DrugB}} - 1} = 1350.$$

Thus, $s_p^2 = \frac{1300(8)+1350(8)}{18} = 1325.$

Plugging into the formula for *t* gives

$$t = \frac{\overline{x}_{\text{DrugA}} - \overline{x}_{\text{DrugB}}}{\sqrt{\frac{s_p^2}{n_{\text{DrugA}}} + \frac{s_p^2}{n_{\text{DrugB}}}}} = \frac{230 - 210}{\sqrt{\frac{1325}{9} + \frac{1325}{9}}}$$

$$= \frac{20}{17.16} = 1.17.$$

Because the researcher is interested in testing whether either drug would outperform the other one, he would conduct a two-tailed independent samples *t* test. The *p* value for this test (as provided by most statistical packages) is .26. As this *p* value is moderately large, he or she would conclude that, although descriptively Drug B outperformed Drug A, this difference might have occurred just by chance.

Alternatively, if a statistical package was not available to determine the precise *p* value, the observed value of *t* can be compared with the critical value. In this case, as observed in most statistics textbooks, the critical *t* value is 2.12. Because $1.17 < 2.12$, the researcher would reach the same conclusion (as will always be the case) and conclude that the difference between the sample means might have occurred just by chance.

Thus, given the results of his study, the researcher would conclude that both drugs are equally effective. Of course, this study does not conclusively settle this issue, but the results of his study do not provide convincing evidence that either drug is better than the other one.

*Eric R. Stone*

*See also* One-Tailed Test; Power; *p* Value; Student's *t* Test; *t* Test, One Sample; *t* Test, Paired Samples; Two-Tailed Test

**Further Readings**

Boneau, C. A. (1960). The effects of violations of assumptions underlying the *t* test. *Psychological Bulletin, 57,* 49–64.
Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
Howell, D. C. (2007). *Statistical methods for psychology* (6th ed.). Belmont, CA: Thomson Wadsworth.
Kirkpatrick, L. A., & Feeney, B. C. (2009). *A simple guide to SPSS for Version 16.0.* Belmont, CA: Thomson Wadsworth

# *t* Test, One Sample

A one-sample *t* test is a hypothesis test for determining whether the mean of a population is different from some known (test) value. The researcher begins by selecting a sample of observations from the population of interest and estimates the population mean by calculating the mean of the sample. The researcher then compares this sample mean with the test value of interest via the formula

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}},$$

where $\overline{x}$ is the sample mean, $\mu$ is the test value, $s$ is the sample standard deviation, and $n$ is the sample size. This $t$ value can then be used to determine the likelihood that any difference between the sample mean and the test value is real versus a result of chance.

For example, a researcher studying the better-than-average effect might be interested in determining whether students, on average, think they are more athletic than the average student. Thus, the researcher has participants rate on a 1–7 scale (where 1 = below average, 4 = average, 7 = above average) their athletic ability and calculates the sample mean. Because the researcher is not interested in the sample per se but in the population of college students, he or she needs to determine whether any difference between the sample mean and "4" (the test value) is real or caused by chance. The one-sample *t* test will assist in making this determination.

## Uses

The one-sample *t* test is appropriate whenever the researcher wants to determine whether the mean of some population differs from some test value. Note the one-sample *t* test closely resembles the one-sample *z* test. The difference is that the *z* test is used when the standard deviation of the population being studied is known, whereas the *t* test is used when the standard deviation of the population is not known.

An important characteristic of the one-sample *t* test (as well as of the one-sample *z* test) is that there is only one population that is being studied. This differentiates a one-sample test from other types of hypothesis tests (e.g., the independent samples *t* test) where there are two (or more) population parameters to be estimated. For example, a researcher interested in whether the average score of males differs from the average score of females would not use a one-sample *t* test, because there are two population parameters (the mean score of both males and females) to be estimated in that situation.

The one-sample *t* test can be used whenever the researcher is interested in determining whether the population mean differs from some specific value (the test value). This test value is generally determined in one of two ways. First, as in the previous example, the test value is chosen to reflect some value of theoretical interest, in that case, the "average" perceived ability. In the second situation, the test value reflects a known population value (or one estimated with a great deal of precision). For example, assume a medical researcher is interested in determining whether a medication affects the number of hours a night that a person sleeps. He finds information from the National Sleep Foundation saying that the average working American adult sleeps 6 hours and 40 minutes (400 minutes) on a work night. He then samples 100 working Americans who take the medication and records how long each of these people sleep on a work night. The mean number of hours slept from this sample could then be compared with 400 minutes via a one-sample *t* test. Note, however, these are

just example uses for the one-sample *t* test. In any situation where the researcher wants to test the mean of a sample against some particular value, a one-sample test is appropriate.

### Interpreting the *t* Value

The result of a one-sample *t* test is a *t* value. This *t* value is compared with the *t*-distribution that would occur if the population mean were equal to the test value. This possibility (that the population mean is equal to the test value) is known as the *null hypothesis*. The logic that underlies the one-sample *t* test is to compare the *observed* value of *t* (that is, the value of *t* calculated from the one-sample *t* test) to the *t*-distribution that would occur if the null hypothesis were true. If the observed value of *t* is in keeping with this distribution, then the conclusion is that the null hypothesis might be true (i.e., the population mean is equal to the test value). If the observed value of *t* is not in keeping with this distribution, then the conclusion is that the null hypothesis is *not* true (i.e., the population mean is not equal to the test value).

The precise form of the *t*-distribution is determined by degrees of freedom (*df*), which in turn is determined by the sample size (*n;* specifically, $df = n - 1$). Basically, the larger the sample size, the better one is able to estimate the standard deviation of the population, so the *t*-distribution most closely resembles the *z* distribution with a large sample size. With a smaller sample size, the *t*-distribution is spread out somewhat to reflect the uncertainty in the estimate of the population standard deviation.

The *t*-distribution for 100 *df* is presented in Figure 1. Intuitively, the observed value of *t* reflects how far from the test value the sample mean is. Because the absolute distance from the mean can be determined by many factors (e.g., the range of possible scores or size of one's sample), this difference is "standardized" by $s/\sqrt{n}$, which is often referred to as the standard error of the mean. Thus, the value of *t* indicates how far the observed mean is from the test value, in standardized form. The *t*-distribution (such as the one presented in Figure 1) provides the range of possible *t* values that would reasonably be expected to be found, *assuming the null hypothesis is true* (i.e., assuming the population mean equals the test value). Put

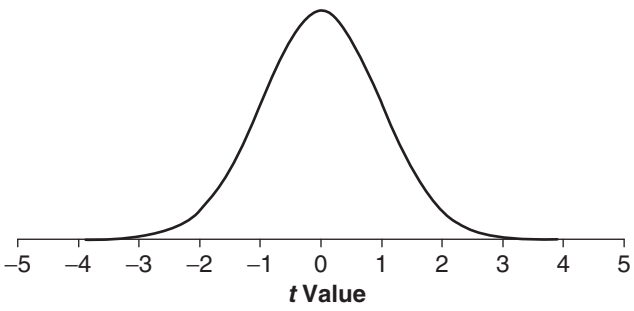



**Figure 1** *t*-Distribution with 100 *df*

differently, even if the population mean equals the test value, the *sample* mean will likely deviate from the test value to some extent just by chance. The *t*-distribution describes precisely how far the sample mean would deviate from the test value just by chance.

To determine whether the population mean could be equal to the test value, then, one compares the *observed t* value with the *t*-distribution constructed under the assumption that the population mean equals the test value. Assume, for example, that the observed *t* value is .50. As can be observed in Figure 1, there is a high probability of obtaining this *t* value if the population mean equals the test value. If, however, the observed *t* value were 4.0, then there is a near-zero probability of obtaining this *t* value if the population mean equals the test value. Thus, in the latter situation, the researcher would conclude that the population mean must *not* equal the test value.

In the previous examples, it was relatively straightforward to tell visually whether the observed value of *t* was in keeping with the *t*-distribution. A more precise indication is provided by a *p* value, which provides the likelihood of obtaining the observed *t* value by chance (i.e., assuming the null hypothesis is true). If the *p* value is low, then the chance of having obtained one's results just by chance (i.e., if the population mean really equals the test value) is low, and the researcher concludes that the population mean is likely not equal to the test value. In this situation, the result of the *t* test is often described as "statistically significant," because it indicates that any difference between the sample mean and test value is real and not just a result of chance. Conversely, if

the *p* value is high, then the likelihood of having obtained one's results just by chance is high, and the researcher concludes that the population mean might be equal to the test value. In many disciplines, .05 is used as a cutoff to determine whether the *p* value is low or high. Using the cutoff of .05 ensures that—of all times when the population mean equals the test value—the researcher will only erroneously conclude that the population mean is *not* equal to the test value 5% of the time. In statistical terms, the researcher will only make a Type I error 5% of the time when the null hypothesis is true.

The exact *p* value is provided by any standard statistical program as a result of conducting a one-sample *t* test. If, however, the one-sample *t* test is conducted by hand, then calculation of the *p* value is not straightforward. Thus, the observed *t* value is sometimes compared with a "critical" value of *t*. This critical *t* value is the value of *t* that would produce a *p* value of .05. Thus, *t* values greater than the critical value of *t* produce *p* values less than .05 and lead to rejection of the null hypothesis (concluding that the population mean is not equal to the test value), and *t* values less than the critical value of *t* produce *p* values greater than .05 and lead to acceptance of the null hypothesis (concluding that the population mean equals the test value). Critical values of *t* can be found in many statistical textbooks.

## One-Tailed Versus Two-Tailed Tests

The exact method for calculating the *p* value (and critical value of *t*) depends on whether one is conducting a directional (one-tailed) test or a nondirectional (two-tailed) test. For a nondirectional test, the researcher is testing whether the population mean is either greater than or less than the test value. In the previous medical example, for instance, the researcher was interested in determining whether the number of hours of sleep either increased or decreased with medication. Thus, the researcher would perform a nondirectional test. If the sample mean was much greater than the test value and provided a large positive value of *t* (see the first equation in this entry), then the researcher would conclude that the medicine increased sleep. Conversely, if the sample mean was much less than the test value and provided a large negative value of *t*, then the researcher would conclude that the medicine decreased sleep. Because the researcher is willing to conclude that there's an effect in either direction (i.e., on either side of the *t*-distribution), this test is referred to as a two-tailed test.

In contrast, with a directional test, the researcher is testing on only one side of the *t*-distribution. For example, in the previous example of the better-than-average effect, the researcher would likely test whether the average student thinks he or she is more athletic than average. Thus, the researcher would perform a directional test, predicting that the population mean was greater than the test value (4 = average, in this example). In this situation, the researcher is only interested in positive values of *t*. If the sample mean was much greater than 4, producing a large positive value of *t*, then she would conclude that the average student thinks he or she is more athletic than average. Otherwise (with a small positive *t* value, or a negative *t* value of any size), she would conclude that people do *not* think they are better than average in athletic ability. Because the researcher is only willing to conclude that there's an effect on one side of the distribution, this test is referred to as a one-tailed test.

The choice of whether to perform a one-tailed or two-tailed test influences the resulting *p* value. As long as the results come out in the predicted direction, the one-tailed test will produce a smaller *p* value (and, equivalently, a smaller critical value of *t*) than will a two-tailed test, thus making it easier to find statistically significant results. However, it is important to emphasize that the decision to conduct a one-tailed test must be made for theoretical reasons and prior to an examination of the data. If, for example, the researcher predicted that the population mean would be greater than 4 and found that the sample mean was 2.8, the researcher must conclude that the hypothesis was not supported, rather than switch to a two-tailed test. For this reason, it is sometimes recommended that researchers typically decide in advance to conduct two-tailed tests to allow for conclusions on either side of the *t*-distribution.

## Assumptions

The three primary requirements for the one-sample *t* test to produce accurate results are unbiased sampling, independent observations, and normality.

The first requirement is unbiased sampling from the population of interest. Note that this requirement must be met for the results of the study to be meaningful descriptively as well as inferentially. Consider the researcher interested in testing whether students on average think they are more athletic than the average student. If she only samples students (say, student athletes) who *are* more athletic than average, then she will conclude that students think they are more athletic than average solely because of the biased sampling of participants. Similarly, in the medical researcher's study, if the Americans taking the medication are somehow systematically different from the average American (for example, having more stressful jobs that reduce the amount of sleep at night), then the medical researcher might find that the participants in his study sleep less than the typical American even if there is actually no effect of the medicine. In short, it is crucial to ensure that the sample constructed for one's study is representative of the population of interest.

A second assumption of the one-sample *t* test is independent observations. Observations in a study are *independent* to the extent that each observation is not predictive of another observation in the study. For example, assume that in the previous hypothetical sleep study, the researcher decided to study five members from the same family for ease of data collection. These values would probably *not* be independent, because each family member is likely to require similar amounts of sleep.

The third assumption of the one-sample *t* test is that the sample was drawn from a population that follows a normal distribution. Partly because of the Central Limit Theorem, however, this assumption can typically be violated as long as the sample size is reasonably large. "Reasonably large" is often defined as $n > 30$; however, as long as the departure from normality is not too severe, considerably smaller sample sizes are acceptable.

## Power

As with any hypothesis test, even when the population mean is different from the test value, the one-sample *t* test might not detect this difference (i.e., the results might not be statistically significant). The ability to find statistically significant results when there *is* a real difference between the population mean and the test value is referred to as the *power* of the test. It is beneficial to conduct one's study with as high a power as possible to provide the greatest chance of being able to detect a difference if one exists.

Given a certain alpha level and choice of one-tailed versus two-tailed test, three factors influence the power of the one-sample *t* test: the difference between the population mean and test value (larger differences = greater power), the amount of variability in the population (smaller variability = greater power), and the size of the sample (larger sample size = greater power). In practice, it is typically easiest to control the size of the sample, so researchers usually try to conduct their studies with as large a sample as possible. Jacob Cohen provides statistical tables that can be used to determine one's power and to determine how large a sample size is needed to provide a given power.

## Example Problem

A used car dealer is trying to determine at what price to sell a 2000 Honda Civic. He thinks that most of his customers overestimate the typical retail price of these cars (the suggested retail price is approximately $8,000, according to the Kelley Blue Book). If he is correct and his customers do overestimate the typical retail price, then he figures he can raise the price of these cars with minimal impact on his selling rate. To determine whether his belief is correct, he asks his next 16 customers to estimate the suggested retail price of the Honda Civic. Their estimates are below (in thousands):

10, 7, 9, 15, 6, 10, 5, 12, 11, 9, 4, 16, 5, 9, 6, 10

The first step in testing his hypothesis is to calculate the sample mean, which is

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{144}{16} = 9.0.$$

Just based on the sample mean, the used car dealer's prediction is confirmed—the average prediction ($9,000) was greater than the suggested retail price ($8,000). However, this difference could have occurred just by chance. To examine this issue, the car dealer would conduct a

one-sample *t* test. The formula for the one-sample *t* test is

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}}.$$

The next step is to calculate the standard deviation ($s$). The formula for the standard deviation is

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}}$$
$$= \sqrt{\frac{(10-9)^2 + (7-9)^2 + \ldots}{15}}$$
$$= \sqrt{\frac{180}{15}} = 3.46.$$

Plugging into the formula for *t* gives

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}} = \frac{9-8}{3.46/\sqrt{16}} = \frac{1}{.865} = 1.16.$$

Because the used car dealer is interested in determining whether people overestimate the actual price, he would conduct a one-tailed one-sample *t* test. The *p* value for this test (provided by most statistical packages) is .12. Because this *p* value is moderately large, he would conclude that, although the average prediction in his sample was greater than $8,000, this difference might well have occurred just by chance.

Alternatively, if a statistical package was not available to determine the precise *p* value, the observed value of *t* can be compared with the critical *t* value. In this case, as found in most statistics textbooks, the critical *t* value is 1.75. Because $1.16 < 1.75$, the used car dealer would reach the same conclusion (as will always be the case) and establish that the difference between the sample mean ($9,000) and the test value ($8,000) might have occurred just by chance.

Thus, given the results of his study, the used car dealer would conclude that people do *not* overestimate the price of 2000 Honda Civics. Of course, this study does not conclusively settle this issue, but the results of his study do not provide convincing evidence that people overestimate these prices.

*Eric R. Stone*

***See also*** One-Tailed Test; Power; *p* Value; Student's *t* Test; *t* Test, Independent Samples; *t* Test, Paired Samples; Two-Tailed Test

**Further Readings**

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Howell, D. C. (2007). *Statistical methods for psychology* (6th ed.). Belmont, CA: Thomson Wadsworth.

Kirkpatrick, L. A., & Feeney, B. C. (2009). *A simple guide to SPSS for Version 16.0*. Belmont, CA: Thomson Wadsworth.

Rosenthal, R., & Rosnow, R. L. (2008). *Essentials of behavioral research: Methods and data analysis*. New York: McGraw-Hill.

## *t* Test, Paired Samples

A paired samples *t* test is a hypothesis test for determining whether the population means of two dependent groups are the same. The researcher begins by selecting a sample of paired observations from the two groups. Thus, each observation in each group is paired (matched) with another observation from the other group. The researcher then calculates the difference between each of these paired observations and conducts a one-sample *t* test on these difference scores via the formula

$$t = \frac{\overline{x}_d}{s_d/\sqrt{n_d}},$$

where $\overline{x}_d$ is the sample mean difference score, $s_d$ is the standard deviation of the sample difference scores, and $n_d$ is the number of paired observations in the sample (i.e., the number of difference scores). This *t* value can then be used to determine the likelihood that any difference between the two sample means is real versus a result of chance.

For example, assume a researcher is interested in the effect of stress on students' performance. She has each of her participants take part in both a "stress" and a "no stress" condition and compares how each participant performs in the two conditions. Because the researcher is not interested in the sample comparisons per se but in all students, she needs to determine whether any

difference between the two conditions she finds in her sample is real or caused by chance. The paired samples *t* test will assist her in making this determination.

Note that the formula given above tests whether the population means are different (i.e., whether the population mean difference score equals zero). It is also possible to use the paired samples *t* test to examine whether the difference between the population means is greater than a constant c (rather than 0). In this situation, the numerator of the formula for *t* changes to $(\overline{x}_d - c)$. In most cases, however, the paired samples *t* test is used to determine whether the population means are equal, and the rest of the discussion is confined to that situation.

## Uses

The paired samples *t* test is appropriate when the researcher wants to know whether the population means of two dependent groups are the same or different. This test is similar to the independent samples *t* test, which also tests whether the population means of two groups are the same or different. The difference between the two tests involves whether the observations in the two groups are paired with each other.

There are multiple reasons why the observations in the two groups might be dependent (paired). The most straightforward situation is when participants take part in both conditions, which is sometimes referred to as a within-subjects design. The stress scenario discussed previously was an example of such a situation; each participant took part in both the stress and no-stress conditions. However, when the observations are paired in any manner, a paired *t* test is appropriate. For example, assume a political strategist is testing the effectiveness of two potential television advertisements. He takes pairs of states with similar voting records and runs one of the two advertisements in one state and the other advertisement in the other state, examining voter preference after seeing the advertisement. For example, New Jersey and Delaware might be considered one pair, Mississippi and Alabama a second pair, and so on.

The advantage of this procedure over randomly assigning each state (participant, etc.) to a condition is that the responses to similar states can be compared. This procedure serves to reduce the amount of variability in the data set, because differences within the states (participants, etc.) are controlled.

## Interpreting the *t* Value

The result of a paired samples *t* test is a *t* value. This *t* value is compared with the *t*-distribution that would occur if the two population group means were equal (i.e., if the mean population difference score was zero). This possibility (that the two population group means are equal) is known as the null hypothesis. The logic that underlies the paired *t* test is to compare the *observed* value of *t* (that is, the value of *t* calculated from the paired *t* test) with the *t*-distribution that would occur if the null hypothesis were true. If the observed value of *t* is in keeping with this distribution, then it can be concluded that the null hypothesis might be true (i.e., the population means are equal). If the observed value of *t* is not in keeping with this distribution, then it can be concluded that the null hypothesis is *not* true (i.e., the population means are not equal).

The precise form of the *t*-distribution is determined by degrees of freedom (*df*), which in turn is determined by the number of paired observations ($n_d$; specifically, $df = n_d - 1$). Basically, the larger the number of paired observations, the better one can estimate the standard deviation of the difference scores, so the *t*-distribution most closely resembles the *z* distribution with a larger number of difference scores.

The *t*-distribution for 100 *df* is presented in Figure 1. Intuitively, the observed value of *t* reflects how different from each other the sample means are. Because the absolute distance between the means can be determined by many factors (e.g., the range of possible scores and size of one's sample), this difference is "standardized" by $s_d/\sqrt{n_d}$, which is the standard error of the mean difference score. Thus, the value of *t* indicates how far apart the sample means are, in standardized form. The *t*-distribution (such as the one presented in Figure 1) provides the range of possible *t* values that would reasonably be expected to be found, *assuming the null hypothesis is true* (i.e., assuming the population means are equal). Put differently, even if the population means are equal, the *sample* means will likely be different just by chance. The

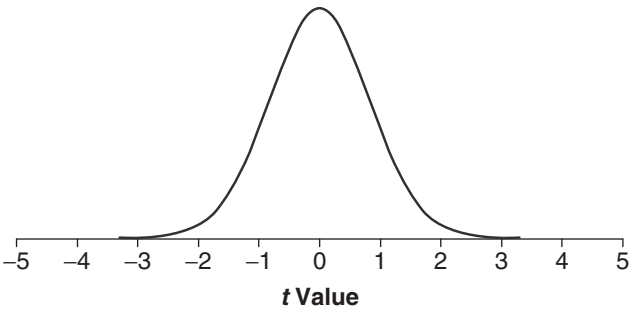



**Figure 1** *t*-Distribution with 100 *df*

*t*-distribution describes precisely how different the sample means would be just by chance.

To determine whether the population means could be equal, then, one compares the *observed t* value with the *t*-distribution constructed under the assumption that the population means are equal. Assume, for example, that the observed *t* value is .50. As can be observed in Figure 1, there is a high probability of obtaining this *t* value if the population means are equal. If, however, the observed *t* value were 4.0, then there is a near-zero probability of obtaining this *t* value if the population means are equal. Thus, in that latter situation, the researcher would conclude that the population means are *not* equal.

In the previous examples, it was relatively straightforward to tell visually whether the observed value of *t* was in keeping with the *t*-distribution. A more precise indication is provided by a *p* value, which provides the likelihood of obtaining the observed *t* value by chance (i.e., assuming the null hypothesis is true). If the *p* value is low, then the chance of having obtained one's results just by chance (i.e., if the population means really are equal) is low, and the researcher concludes that the population means are likely not equal. In this situation, the result of the *t* test is often described as "statistically significant," because it indicates that any difference between the sample means is real and not just a result of chance. Conversely, if the *p* value is high, then the likelihood of having obtained one's results just by chance is high, and the researcher concludes that the population means might be equal. In many disciplines, .05 is used as a cutoff to determine whether the *p* value is low or high. Using the cutoff of .05 ensures that—of all times when the population means are equal—the researcher will only erroneously conclude that the population means are *not* equal 5% of the time. In statistical terms, the researcher will only make a Type I error 5% of the time when the null hypothesis is true.

The exact *p* value is provided by any standard statistical program as a result of conducting a paired *t* test. If, however, the paired *t* test is conducted by hand, calculation of the *p* value is not straightforward. Thus, the observed *t* value is sometimes compared with a "critical" value of *t*. This critical *t* value is the value of *t* that would produce a *p* value of .05. Thus, *t* values greater than the critical value of *t* produce *p* values less than .05 and lead to rejection of the null hypothesis (concluding that the population means are not equal), and *t* values less than the critical value of *t* produce *p* values greater than .05 and lead to acceptance of the null hypothesis (concluding that the population means are equal). Critical values of *t* can be found in many statistical textbooks.

## One-Tailed Versus Two-Tailed Tests

The exact method for calculating the *p* value (and critical value of *t*) depends on whether one is conducting a directional (one-tailed) test or a nondirectional (two-tailed) test. For a nondirectional test, the researcher is testing whether the population means are different (i.e., whether either mean is greater than the other one). In the previous stress example, for instance, the researcher might be interested in determining whether stress either increases or decreases performance. In this case, the researcher would perform a nondirectional test. If the sample mean of the no-stress group was much greater than the sample mean of the stress group, the researcher would conclude that stress decreased performance. Conversely, if the sample mean of the stress group was much greater than the sample mean of the no-stress group, the researcher would conclude that stress increased performance.

To see how this works mathematically, consider Figure 1. We will call the no-stress group Group 1 and the stress group Group 2 (the choice is arbitrary). In that case, when stress decreases performance in the sample ($\overline{x}_{\text{nostress}} > \overline{x}_{\text{stress}}$), the researcher will obtain a positive value of *t,* and

when stress increases performance in the sample ($\overline{x}_{\text{nostress}} < \overline{x}_{\text{stress}}$), the researcher will obtain a negative value of *t*. In either case, as long as the difference between the sample means is sufficiently large, the researcher will obtain a large (positive or negative) value of *t,* leading to rejection of the null hypothesis and the conclusion that there really is a difference between the groups. Because the researcher is willing to conclude that there is an effect in either direction (i.e., on either side of the *t*-distribution), this test is referred to as a two-tailed test.

In contrast, with a directional test, the researcher is testing on only one side of the *t*-distribution. For instance, in the previous example, the researcher might be interested in testing (only) whether stress decreases performance. In this case, the researcher would perform a directional test, predicting that the population group mean for the no-stress group would be greater than the population group mean for the stress group. In this situation, the researcher is only interested in positive values of *t*. If the sample mean for the no-stress group was much greater than the sample mean for the stress group, producing a large positive value of *t,* the researcher would conclude that stress decreases performance. Otherwise (with a small positive *t* value or a negative *t* value of any size), the researcher would conclude that stress does *not* decrease performance. Because the researcher is only willing to conclude that there is an effect on one side of the distribution, this test is referred to as a one-tailed test.

The choice of whether to perform a one-tailed or two-tailed test influences the resulting *p* value. As long as the results come out in the predicted direction, the one-tailed test will produce a smaller *p* value (and, equivalently, a smaller critical value of *t*) than will a two-tailed test, thus making it easier to find statistically significant results. However, it is important to emphasize that the decision to conduct a one-tailed test must be made for theoretical reasons and prior to an examination of the data. If, for example, the researcher predicted that stress would decrease performance and found that the sample mean of the stress group was greater than the sample mean of the no-stress group, the researcher must conclude that the hypothesis was not supported, rather than switch to a two-tailed test. For this reason, it is sometimes recommended that researchers typically decide in advance to conduct two-tailed tests to allow for conclusions on either side of the *t*-distribution.

## Assumptions

The two primary requirements for the paired samples *t* test to produce accurate results are independent observations and normality.

The first assumption of the paired *t* test is independent observations. Observations in a study are *independent* to the extent that each observation is not predictive of another observation in the study. Note that it is the scores within each group that must be independent of each other (or more precisely, the difference scores that need to be independent). For example, assume that in the previous hypothetical stress study, the researcher tested multiple members of one family. It is plausible that the difference scores would not be independent, because it is likely that stress would influence the family members in a near identical way.

The second assumption of the paired *t* test is that the population distribution of difference scores is normal. Partly because of the Central Limit Theorem, however, this assumption can typically be violated as long as the sample size is reasonably large. "Reasonably large" is often defined as $n > 30$; however, as long as the departure from normality is not too severe, considerably smaller sample sizes are acceptable.

## Power

As with any hypothesis test, even when the population group means are different from one another, the paired samples *t* test might not detect this difference (i.e., the results might not be statistically significant). The ability to find statistically significant results when there *is* a real difference between the population means is referred to as the *power* of the test. It is beneficial to conduct one's study with as high a power as possible, to provide the greatest chance of being able to detect a difference if one exists.

Given a certain alpha level and choice of one-tailed versus a two-tailed test, there are three factors that influence the power of the paired *t* test: the difference between the population means

(larger differences = greater power), the amount of variability in the population difference scores (smaller variability = greater power), and the size of the samples (larger sample size = greater power). Jacob Cohen provides statistical tables that can be used to determine one's power and to determine how large a sample size is needed to provide a given power.

As discussed previously, the process of "pairing" (whether via a within-subjects design or some other procedure) serves to reduce the variability in the study (because what is relevant is the variability in the difference scores, not the variability within each of the groups). Thus, the power of the paired *t* test is typically greater than that of other tests that do not use this pairing procedure (such as the independent samples *t* test).

## Example Problem

A nutritionist is interested in determining whether eating breakfast improves cognitive performance. He enlists the participation of 16 4th-graders and has them take an IQ test on one day after eating breakfast and on another day when they have not eaten breakfast, counterbalancing appropriately. The scores are in Table 1, under the columns labeled "Breakfast" and "No Breakfast."

**Table 1** *t* Test, Paired Samples

| *Participant* | *Breakfast* | *No Breakfast* | *Difference* |
|---|---|---|---|
| 1 | 95 | 88 | +7 |
| 2 | 112 | 116 | –4 |
| 3 | 98 | 99 | –1 |
| 4 | 102 | 93 | +9 |
| 5 | 81 | 75 | +6 |
| 6 | 127 | 120 | +7 |
| 7 | 110 | 113 | –3 |
| 8 | 94 | 89 | +5 |
| 9 | 103 | 96 | +7 |
| 10 | 89 | 84 | +5 |
| 11 | 133 | 129 | +4 |
| 12 | 109 | 113 | –4 |
| 13 | 99 | 91 | +8 |
| 14 | 104 | 107 | –3 |
| 15 | 91 | 85 | +6 |
| 16 | 114 | 115 | –1 |

The first step in testing his hypothesis is to calculate the difference scores for each participant. These are provided in the column labeled "Difference." Note that positive scores indicate better performance in the breakfast condition, and negative scores indicate better performance in the no-breakfast condition.

The next step is to calculate the mean difference score, which is

$$\overline{x}_d = \frac{\sum_{i=1}^{n} x_i}{n_d} = \frac{48}{16} = 3.0.$$

Note that this mean difference score could also be calculated by subtracting the mean no-breakfast score from the mean breakfast score.

Just based on the sample means, it seems that eating breakfast led to higher IQ scores. However, it is possible that this effect occurred just by chance. To examine this issue, the nutritionist would conduct a paired samples *t* test. The formula for the paired samples *t* test is

$$t = \frac{\overline{x}_d}{s_d/\sqrt{n_d}}.$$

The next step is to calculate the standard deviation of the difference scores ($s_d$). The formula for $s_d$ is

$$\begin{aligned} s_d &= \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x}_d)^2}{n_d - 1}} \\ &= \sqrt{\frac{(7-3)^2 + (-4-3)^2 + \cdots}{15}} \\ &= \sqrt{\frac{338}{15}} = 4.747. \end{aligned}$$

Plugging into the formula for *t* gives

$$t = \frac{\overline{x}_d}{s_d/\sqrt{n_d}} = \frac{3}{4.747/\sqrt{16}} = \frac{3}{1.187} = 2.53.$$

Because the nutritionist is interested in testing whether eating breakfast improves cognitive performance, he would conduct a one-tailed paired *t* test. The *p* value for this test (provided by most statistical packages) is .01. Because this *p* value is low, he would conclude that his results did not

occur by chance and that eating breakfast likely does improve cognitive performance.

Alternatively, if a statistical package was not available to determine the precise $p$ value, the observed value of $t$ can be compared with the critical $t$ value. In this case, as found in most statistics textbooks, the critical $t$ value is 1.75. Because $2.53 > 1.75$, the nutritionist would reach the same conclusion (as will always be the case), that the sample mean difference of 3 points did not just occur by chance.

Thus, given the results of his study, the nutritionist would conclude that eating breakfast increases performance on the IQ test. Of course, this study would not conclusively prove that this is the case; however, the likelihood that this difference occurred just by chance is low.

*Eric R. Stone*

***See also*** One-Tailed Test; Power; $p$ value; Student's $t$ test; $t$ Test, One Sample; $t$ Test, Independent Samples; Two-Tailed Test

**Further Readings**

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Howell, D. C. (2007). *Statistical methods for psychology* (6th ed.). Belmont, CA: Thomson Wadsworth.

Kirkpatrick, L. A., & Feeney, B. C. (2009). *A simple guide to SPSS for Version 16.0*. Belmont, CA: Thomson Wadsworth.

Kurtz, K. H. (1965). *Foundations of psychological research*. Boston: Allyn & Bacon.

# Tukey's Honestly Significant Difference (HSD)

The Tukey's honestly significant difference test (Tukey's HSD) is used to test differences among sample means for significance. The Tukey's HSD tests all pairwise differences while controlling the probability of making one or more Type I errors. The Tukey's HSD test is one of several tests designed for this purpose and fully controls this Type I error rate. Other tests such as the Newman–Keuls lead to an inflated Type I error rate in some situations. This entry describes how to conduct and interpret the Tukey's HSD test.

## Philosophy

It is rare for any two experimental treatments to have identical effects. For example, it is implausible that two drug treatments could produce the same relief from depression if measured to 100 decimal places. As a result, the role of inferential statistics in these situations is not to reject the null hypothesis of no difference, because that hypothesis is false on its face. Instead, it is to determine whether a confident statement can be made about the direction of the difference. Depending on the results of the inferential test, a researcher might be able to state with confidence the direction of the difference, might have a hint about the direction, or might have little or no information about the direction. Clearly, an all-or-nothing decision rule in which one either rejects or fails to reject the null hypothesis is not consistent with this approach. Instead, the probability value obtained in the inferential test is used to aid in the assessment of the confidence one should have in the direction of the difference.

Because some treatments of inferential statistics take the more traditional approach of testing whether a difference is exactly zero, it is important to show the correspondence between the approach taken here and the traditional approach. In the traditional approach, a Type I error is defined as rejecting a true null hypothesis. Here, a Type I error is defined as making a confident claim about the direction of the difference when there is either no difference or when the difference is in the opposite direction of the claimed difference. The Type I error rate in this context depends on the size of the true difference between means: If there is a large difference between the means, then the probability of getting the direction wrong is smaller than when there is a small difference between means. In computing the Type I error rate, the conservative approach is to assume that the true difference is zero. Although this might not often represent reality, it is the best way to ensure that the Type I error rate is controlled and is adopted here.

**Table 1** The Six Comparisons Among Four Treatment Conditions

| *Comparison* | *Conditions* | *Compared* |
|---|---|---|
| 1 | 1 | 2 |
| 2 | 1 | 3 |
| 3 | 1 | 4 |
| 4 | 2 | 3 |
| 5 | 2 | 4 |
| 6 | 3 | 4 |

## The Problem of Multiple Comparisons

If a researcher compared the means of four treatment conditions, there would be six pairwise comparisons. These comparisons are shown in Table 1.

It is important to distinguish between the following two error probabilities: 1) the probability that any single comparison results in a Type I error and 2) the probability that one or more comparisons result in a Type I error. The former probability is referred to as the per-comparison error rate; the latter probability is referred to as the family-wise error rate.

Because there are six comparisons in this example, the researcher has six chances to make a Type I error. Naturally, the more chances one has of making a Type I error, the higher the probability that at least one Type I error will be made. The probability is highest when the comparisons are independent. When independent, the probability of at least one Type I error can be computed using the following formula:

$$p_f = 1 - (1 - p_c)^c$$

where $p_f$ is the family-wise error rate and is the probability of making one or more Type I errors, $p_c$ is the per-comparison error rate and is the probability of making a Type I error in a single comparison, and $c$ is the number of comparisons. For example, if the probability of making a Type I error with a single comparison were .05 and six comparisons were made, then the probability of one or more Type I errors would be

$$p_f = 1 - (1 - 0.05)^6 = 0.265.$$

This formula overstates the family-wise error rate because comparisons of pairs of means are not independent. In this example, consider whether a finding that Mean 1 is greater than Mean 2 would have any bearing on the comparison of Mean 1 and Mean 3. It turns out that if Mean 1 is larger than Mean 2, then it is likely that Mean 1 is larger than Mean 3. This can be observed from Table 2, which lists all possible orders of means. Of the six orders, Mean 1 is higher than Mean 2 in three orders (orders 1, 2, and 5). The critical question is, What happens when only these three orders are considered? Looking only at these three orders, it can be observed that Mean 1 is larger than Mean 2 in two of the three orders (orders 1 and 2). Therefore, the probability that Mean 1 will be greater than Mean 3 given that Mean 1 is greater than Mean 2 is .67. Clearly, the comparison of Mean 1 with Mean 2 is not independent of the comparison of Mean 1 with Mean 3.

Because these comparisons are not independent, the calculation that $p_f = .265$ overstates $p_f$. Although the details of how to calculate $p_f$ for this example are beyond the scope of this entry, $p_f$ is about .12. Therefore, although $p_f$ is considerably below the value of .265 that is based on the assumption that the comparisons are independent, it is still considerably above the per-comparison rate of .05.

The degree to which the family-wise error rate is higher than the per-comparison rate depends on the number of comparisons, which, in turn, depends on the number of conditions in the experiment. The number of comparisons can be computed from the following formula, where $c$ is the

**Table 2** Demonstration of Nonindependence of Comparisons

| *Order* | *Largest* | *Middle* | *Smallest* | *Mean 1 > Mean 2?* | *Mean 1 > Mean 3?* |
|---|---|---|---|---|---|
| 1 | 1 | 2 | 3 | Yes | Yes |
| 2 | 1 | 3 | 2 | Yes | Yes |
| 3 | 2 | 1 | 3 | No | |
| 4 | 2 | 3 | 1 | No | |
| 5 | 3 | 1 | 2 | Yes | No |
| 6 | 3 | 2 | 1 | No | |

number of comparisons and $k$ is the number of conditions:

$$c = \frac{k(k-1)}{2}$$

As shown in the Table 3, the number of comparisons increases rapidly with the number of conditions.

The family-wise error rate for 10 conditions is about 0.80 and the rate for 20 conditions is about 0.90. It is obvious that the family-wise error rate can be extremely high if steps are not taken to control it.

One might be tempted simply to compute a $t$ test between the largest and smallest means and, because only one test is computed, interpret it in the usual way. However, such a test would be greatly biased because the to-be-compared means would have been chosen after the data were examined. As a result, this procedure is tantamount to testing all pairwise differences and thus would result in the same inflated family-wise error rate as testing all pairwise comparisons.

## Statistical Basis

The Tukey's HSD test controls the family-wise Type I error rate. The basic approach is to compute the difference between the largest and smallest means and determine the probability of obtaining a difference that is big or bigger if all the population means were equal. Because the range is defined as the difference between largest and smallest values, the probability in question is the probability of finding means with a range as large or larger than the range of means in the sample data.

**Table 3** Number of Comparisons as a Function of Number of Conditions

| *Conditions* | *Comparisons* |
|---|---|
| 2 | 1 |
| 3 | 3 |
| 4 | 6 |
| 5 | 10 |
| 10 | 45 |
| 20 | 190 |

The significance test is based on the studentized range distribution. The studentized range distribution is similar to the Student's $t$-distribution but differs in that it takes into account the number of means being compared. The formula for a $t$ test comparing two independent means is given below:

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{2MSE}{n}}},$$

where $\overline{x}_1$ is the sample mean for group 1, $\overline{x}_2$ is the sample mean for group 2, $n$ is the sample size for each group, and $MSE$ is the estimated population variance computed by averaging the sample variances in the two groups. The probability value for the $t$ is computed based on the degrees of freedom which is equal to $N - 2$, where $N$ is the total number of observations.

The Tukey's HSD tests the difference between each pair of means using the value of the studentized $t$ ($t_s$) computed using the following formula:

$$t_s = \frac{\overline{x}_i - \overline{x}_j}{\sqrt{\frac{MSE}{n}}},$$

where $\overline{x}_i$ is the mean of group $i$, $\overline{x}_j$ is the mean of group $j$, and $MSE$ is the estimated population variance computed by averaging the sample variances of all the groups. The probability value for $t_s$ is computed based on two parameters: the degrees of freedom ($df$) and the number of conditions in the experiment ($k$). The degrees of freedom is equal to $N - k$, where $N$ is the total number of observations. Probability values for the studentized range distribution are calculated by some but not all statistical packages. An online calculator for the studentized range is available at http://onlinestatbook.com/analysis_lab/studentized_range_dist.html.

There are two differences between the formulas for $t$ and $t_s$: the formula for $t$ compares the means of the two conditions in an experiment and therefore there is only one value of $t$. The $t_s$ formula compares $\overline{x}_i$ and $\overline{x}_j$ for all combinations of $i$ and $j$ except for the trivial cases in which $i$ equals $j$. Therefore, if there are three conditions, the values of $i$ and $j$ are 1, 2; 1, 3; and 2, 3 for the three comparisons among the means. These

represent comparisons of Mean 1 with Mean 2, Mean 1 with Mean 3, and Mean 2 with Mean 3, respectively. Second, notice there is a "2" in the denominator for $t$ but not for $t_s$. This difference is more or less a historical accident and has no practical consequence because the definition of the studentized range assumes the computation does not involve a "2" in the denominator. The studentized range could have been defined in such a way that the formula for $t_s$ did include this "2."

The $t$-distribution is equal to the studentized range distribution when $k = 2$ except for whether there is a "2" denominator. Because the calculation of probabilities takes this difference into account, the $p$ values are always the same. For example, for a difference between means of 12, a sample size (per group) of 10, and an *MSE* of 100, the values of $t$ and $t_s$ are 2.68 and 3.79, respectively. The $p$ value for a $t$ of 2.68 and a $t_s$ of 3.79 are both equal to .0137.

## Assumptions

The Tukey's HSD test assumes that (a) all observations are sampled randomly and independently, (b) the $k$ populations are distributed normally, and (c) the $k$ populations all have the same variance. This last assumption is called the assumption of homogeneity of variance.

## Displaying the Results

There are a variety of ways the results of the Tukey's HSD test can be displayed. A common approach is to use underlining to indicate which means are significantly different from each other and which means are not. In the example shown below, means that are not significantly different share a line.

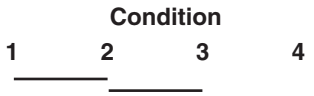



**Figure 1** Commonly Used Method to Indicate Which Means Are Significantly Different

**Table 4** Probability Values and 95% Confidence Intervals

| | *1* | *2* | *3* | *4* |
|---|---|---|---|---|
| 1 | *X* | –0.88 to 3.55 | 2.28 to 5.72 | 7.95 to 12.38 |
| 2 | .358 | *X* | –0.05 to 4.38 | 6.62 to 11.05 |
| 3 | .001 | .057 | *X* | 4.45 to 8.88 |
| 4 | < .001 | < .001 | < .001 | *X* |

Because there is a line between Conditions 1 and 2, these conditions are not significantly different. Similarly, there is a line between Conditions 2 and 3, which indicates these conditions are not significantly different. As there is no line between Conditions 1 and 3, Conditions 1 and 4, Conditions 2 and 4, and Conditions 3 and 4, these four differences are all significant. This approach is not recommended, however, because it follows the all-or-none rejection of the null hypothesis philosophy.

A better method for displaying the results is to give the probability values and confidence intervals for each comparison in a matrix as shown in Table 4.

The upper triangle contains the 95% confidence intervals for each difference between means. For example, the entry "$-0.88$ to 3.55" in the first row, second column is the 95% confidence interval for $\overline{x}_2 - \overline{x}_1$. The interval has a lower limit of $-0.88$ and an upper limit of 3.55. Because both positive and negative values are included in the interval, a confident statement about the direction of the difference cannot be made. The lower triangle contains the probability values. The value in the second row and first column is .358. This is the two-tailed probability for the comparison of Means 1 and 2. Consistent with the confidence interval (as it always will be), this result does not allow a confident conclusion about direction.

## Worked Example

Table 5 shows the raw data from a hypothetical experiment in which there are four conditions and six observations in each condition. The data for a given condition is in the column labeled by the number of the condition.

**Table 5** Example Data

| *Condition* | | | |
|---|---|---|---|
| *1* | *2* | *3* | *4* |
| 3 | 4 | 8 | 12 |
| 2 | 5 | 6 | 13 |
| 4 | 4 | 6 | 12 |
| 5 | 5 | 5 | 15 |
| 4 | 4 | 8 | 16 |
| 3 | 7 | 9 | 14 |

**Table 6** Means and Variances as a Function of Condition

| | *Condition* | | | |
|---|---|---|---|---|
| | *1* | *2* | *3* | *4* |
| Mean | 3.500 | 4.8333 | 7.000 | 13.667 |
| Variance | 1.100 | 1.3667 | 2.400 | 2.6667 |

**Table 7** Tests for the Comparisons Among Conditions

| *Condition i* | *Condition j* | *Difference* | $t_s$ |
|---|---|---|---|
| 4 | 1 | 10.167 | 18.15 |
| 4 | 2 | 8.833 | 15.77 |
| 4 | 3 | 6.667 | 11.90 |
| 3 | 1 | 3.500 | 6.25 |
| 3 | 2 | 2.167 | 3.87 |
| 2 | 1 | 1.333 | 2.38 |

The first step in computing the Tukey's HSD is to compute the means and variances for each condition. These are shown in Table 6.

The next step is to compute the *MSE*. Recall that the *MSE* is the mean of the variances. For these data, $MSE = 1.883$. The denominator of the formula for $t_s$ can be computed as follows:

$$\sqrt{\frac{MSE}{n}} = \sqrt{\frac{1.883}{6}} = 0.560$$

To compute $t_s$ for a comparison, the difference between means is divided by the denominator of 0.560. The values of $t_s$ are shown in Table 7.

The probability values can be computed from the calculator referenced in the "Statistical Basis" section. Note that the probability values shown in the "Displaying the Results" section are based on these data.

If a studentized range calculator is not available, the critical values for significance testing can be found in a table of the studentized range distribution. Typically, the rows show the degrees of freedom, the columns show the number of conditions, and each entry contains a critical value. As discussed, it is better to report probability values than whether a difference is significant. Therefore, tables of the studentized range should be used only as a last resort.

## Unequal Sample Sizes

The formula given previously for $t_s$ has $n$, the sample size for each group, in the denominator. Therefore, a different formula is required when the sample sizes are not equal. A good solution is to use what is called the Tukey-Kramer Method. The formula for unequal $n$s is

$$t_s = \frac{\overline{x}_i - \overline{x}_j}{\sqrt{\frac{MSE}{n_h}}},$$

where $n_h$ is the harmonic mean of $n_i$ and $n_j$ and is computed as follows:

$$n_h = \frac{2}{\frac{1}{n_i} + \frac{1}{n_j}}.$$

Therefore, if $n_i = 12$ and $n_j = 16$ then

$$n_h = \frac{2}{\frac{1}{12} + \frac{1}{16}} = 13.71.$$

Notice that the harmonic mean is slightly lower than the arithmetic mean of 14.

When the sample sizes are not equal, *MSE* is computed somewhat differently so as to take into account the fact that some sample variances are based on more observations than others. For each group, the sum of squares error is computed using the numerator for the formula for the variance. That is, for group $i$, the sum of squares error (*SSE*) is

$$SSE_i = \sum (X - \overline{x}_i)^2.$$

The SSE is then the sum of the SSEs for the separate groups. Finally, *MSE* is computed by dividing SSE by the degrees of freedom error (*dfe*). The degrees of freedom error is equal to $N - k$.

## Violations of Assumptions

The Tukey's HSD is relatively robust to violations of the normality assumption and, in general, provides conservative probability values. However, a violation of the assumption of homogeneity of variance can seriously inflate the Type I error rate. There are a variety of approaches to handling heterogeneity of variance including Dunnett's T3 test, Dunnett's C test, and the Games-Howell test. The calculations of these tests and their associated probability values is complex and beyond the scope of this entry. Researchers should use the Tukey's HSD cautiously when there is evidence of heterogeneity of variance.

## Is the Tukey's HSD Test a Post Hoc Test?

Some statistics texts and software packages classify the Tukey's HSD test as a post hoc test and/or as a test performed following a significant analysis of variance. Classifying the Tukey's HSD as post hoc means that the decision to use this test was made after having analyzed or viewed the data. Although this might be true in some instances, there is no reason a researcher cannot (or should not) plan a priori to compare all pairs of means using the Tukey's HSD. It is therefore a misnomer to classify the Tukey's HSD as post hoc.

The Tukey's HSD test is often presented as a follow-up to a significant analysis of variance (ANOVA). The thinking is that if the ANOVA is significant, then the omnibus null hypothesis that all population means are equal can be rejected. The Tukey's HSD is used as a follow-up to determine which means are significantly different from other means. Although this sequential testing scheme is valid, it is not necessary. The Tukey's HSD when used alone controls the family-wise Type I error rate. There is, therefore, no requirement that the Tukey's HSD be conducted following a significant analysis of variance.

Because ANOVA and the Tukey's HSD have different statistical bases, they occasionally lead to different results. That is, it is possible for the ANOVA to fail to be significant while one or more of the Tukey's HSD comparisons is significant. It is also possible for the ANOVA to be significant and none or the Tukey's HSD comparisons to be significant.

It is instructive to consider the costs and benefits of these two procedures when they lead to different results. First, consider the case in which the ANOVA is not significant and a Tukey's HSD comparison is. In this case, the sequential procedure would not lead to any confident conclusion, whereas the Tukey's HSD alone would lead to a confident conclusion about the direction of at least one difference between means. In the case in which the ANOVA is significant and none of the Tukey's HSD comparisons are significant, the sequential procedure would allow the rejection of the omnibus null hypothesis, whereas the Tukey's HSD alone would not result in any confident conclusions. Thus, the potential loss of the sequential procedure is to miss out on a conclusion about a particular pair of means, whereas the potential loss of the Tukey's HSD alone is to miss out on rejecting the omnibus null hypothesis. The former loss would seem more serious than the latter so the recommended procedure is to do the Tukey's HSD alone.

*David Mark Lane*

***See also*** Analysis of Variance (ANOVA); Newman–Keuls Test and Tukey Test; Normality Assumption; Null Hypothesis; Pairwise Comparisons; Post Hoc Comparisons; Power; *p* Value; *t* Test, Independent Samples; Type I Error

### Further Readings

Kirk, R. E. (1968). *Experimental design: Procedures for the behavioral sciences*. Belmont, CA: Thomson Wadsworth

Tukey, J. W. (1991) The philosophy of multiple comparisons. *Statistical Science, 6,* 100–116.

Wilkinson, L., & the Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.

# Two-Tailed Test

A two-tailed test is a statistical procedure used to compare the null hypothesis (that a population parameter is equal to a particular value) against the alternative hypothesis (that the population parameter is *different* from this value). Evidence regarding the null hypothesis is obtained from a test statistic, and the test is said to be "two tailed" because its alternative hypothesis does not specify whether the parameter is greater than or less than the value specified by the null hypothesis. Hence, both large and small values of the test statistic, that is, values on both tails of its distribution, provide evidence against the null hypothesis. This type of test is relevant for situations in which researchers wish to test a null hypothesis, but they do not have a prior belief about the direction of the alternative, a situation that is likely to happen in practice. The term *two-tailed test* is usually reserved for the particular case of one-dimensional hypotheses, even though it might be used more generally.

## Two-Sided Hypothesis Testing

In hypothesis testing, the hypotheses are always statements about a population parameter that partitions the set of possible values that the parameters might take. For example, letting $\mu$ be the parameter for which the hypothesis test is performed, a null hypothesis, which is referred to as $H_0$, might be defined as

$$H_0 : \mu = \mu_0,$$

and its two-sided alternative hypothesis, which is referred to as $H_1$, is defined as

$$H_1 : \mu \neq \mu_0.$$

The alternative hypothesis $H_1$ does not make a statement about whether $\mu$ is greater than $\mu_0$ or less than $\mu_0$, which makes this a two-sided test. The difference between a one-sided test and a two-sided test lies solely in the specification of the alternative hypothesis. As a consequence, whereas a one-sided test specifies in its alternative hypothesis that the parameter is either greater than or less than the value specified in the null hypothesis ($H_1$ is either $\mu > \mu_0$ or $\mu < \mu_0$), in a two-sided test, the direction of the alternative hypothesis is left unspecified.

Evidence for or against the null hypothesis is obtained by means of a test statistic, which is a function of the available data. Just as in the one-sided case, in a two-sided hypothesis test the decision of whether to reject the null hypothesis $H_0$ is based on a test statistic $W(X) = W(X_1, X_2, \ldots, X_N)$ which is a function of a (random) sample $X_1, X_2, \ldots, X_N$ of size $N$ from the population under study. The test specifies a rejection rule that indicates in what situations $H_0$ should be rejected. In a two-sided test, rejection occurs for both large and small values of $W(X)$, whereas in a one-sided test, rejection occurs either for large or small values of the test statistic (but not both) as dictated by the alternative hypothesis. Formally, a two-sided rejection rule is defined as:

Reject $H_0$ if $W(X) < c_1$ or $W(X) > c_2$.

Do not reject $H_0$ if $c_1 \leq W(X) \leq c_2$.

To establish the values of the critical values $c_1$and $c_2$, it is common practice to follow the Neyman–Pearson approach and first choose a significance level $\alpha$. The significance level $\alpha$ of the test is an upper bound to the probability of mistakenly rejecting $H_0$ when $H_0$ is true (probability of Type I error). Once the significance level has been fixed, the constants $c_1$and $c_2$ are chosen so that the probability of rejecting $H_0$ when $H_0$ is true is (at most) equal to the significance level. In other words, $c_1$ and $c_2$ are chosen so that

$$\Pr_{H_0}[W(X) < c_1] + \Pr_{H_0}[W(X) > c_2] \leq \alpha,$$

where $\Pr_{H_0}(z)$ indicates the probability of $z$ computed assuming that the null hypothesis $H_0$ is true.

This still might leave the constants $c_1$ and $c_2$ undetermined, because there might be infinitely many ways in which the sum of these two terms can be made equal to $\alpha$. Thus, the researcher must usually make a decision regarding how to divide the probability $\alpha$ between the two terms, that is, between the two tails of the distribution of $W(X)$ under $H_0$. If the researcher has no prior information regarding the direction of the alternative, then it seems appropriate to divide this total probability

symmetrically between the two tails. That is, the condition $\Pr_{H_0}[W(X) < c_1] = \Pr[W(X) > c_2]$ is imposed, and therefore

$$\Pr_{H_0}[W(X) < c_1] = \Pr_{H_0}[W(X) > c_2] \leq \frac{\alpha}{2}.$$

If the researcher has prior information regarding the population parameter that might affect the alternative hypothesis, then this total probability might be divided asymmetrically between the two tails. However, an asymmetric allocation of $\alpha$ between both tails is not used very often, because in cases when information regarding the direction of the effect under study is available, researchers usually choose a one-sided alternative.

The two-sided rejection rule is easier to construct when the distribution of $W(X)$ under the null hypothesis is symmetric, because in this case, the critical values $c_1$ and $c_2$ are equal in absolute value. In this case, there is only one unknown constant that needs to be established based on the underlying distribution of the test statistic.

## Comparison With One-Sided Test

The difference in the specification of the alternative hypothesis between a one-tailed test and a two-tailed test has important conceptual consequences. As illustrated in the example that follows, using a two-sided test is generally conservative in the sense that it is more difficult to reject the null hypothesis with this test than with the correct one-sided test for a given significance level. This occurs because a more extreme value of the test statistic will be necessary to reject the null hypothesis at the same $\alpha$ significance level with a two-sided test than with a one-sided test, because in the former, the total probability of rejecting $H_0$when it is true (Type I error) is split between both tails of the distribution of $W(X)$.

For example, when the null hypothesis $H_0$ is tested using both an $\alpha$-level one-sided test to the right and an $\alpha$-level two-sided test, and the distribution of the test statistic is continuous, the critical value $c^*$ of the one-sided test is defined by $\Pr_{H_0}[W(X) > c^*] = \alpha$, and the critical values $c_l^{**}$ and $c_u^{**}$ of the two-sided test are defined by $\Pr_{H_0}\left[W(X) < c_l^{**}\right] + \Pr_{H_0}\left[W(X) > c_u^{**}\right] = \alpha$. It is easy to see that in this case $c^* < c_u^{**}$ and values of $W(X)$ exist such that $c^* < W(X) < c_u^{**}$. When this happens, $H_0$ will be rejected with the one-sided test but will not be rejected with the two-sided test.

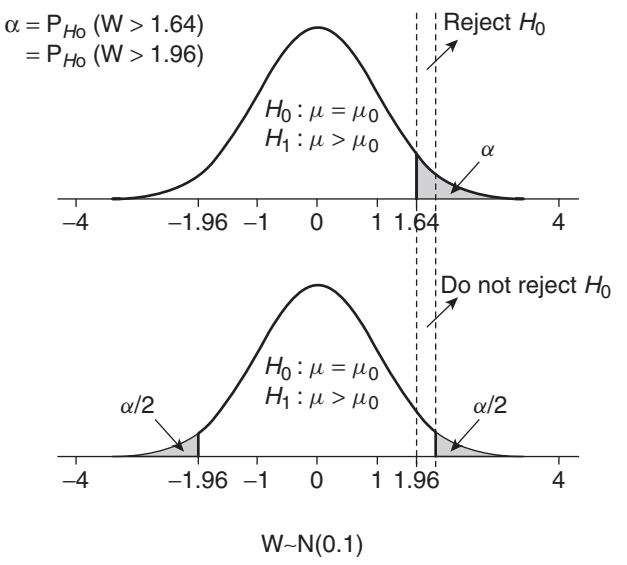



**Figure 1** One-Sided Versus Two-Sided Hypothesis Test Under Normality

This point is further illustrated in Figure 1, where the top panel shows the significance level of a one-sided hypothesis test for the particular case of a normal distribution of the test statistic under $H_0$, and the bottom panel shows a two-sided test with the same significance level and the same test statistic, where the significance level has been split symmetrically across both tails. As can be viewed in the figure, for all values of $W(X)$ between 1.64 and 1.96, the null hypothesis is rejected with a one-sided test but is *not* rejected with a two-sided test. The two-sided test requires a larger value of $W(X)$ to reject $H_0$ than the one-sided test shown in the figure, because the probability of Type I error on the upper tail is forced to be smaller in the two-sided test ($\alpha/2$) than in the one-sided test ($\alpha$). This illustrates how a two-sided test might require a more surprising value of $W(X)$ to reject the null hypothesis than a one-sided test, which makes the two-sided test more conservative.

## A Numerical Example

Imagine a situation in which a researcher is interested in establishing whether two competing math

text books have the effect of increasing the mathematical skills of elementary school students. In particular, the researcher is interested in whether assigning the practice exercises of the books as homework has an effect on math test scores. For this purpose, $N$ students are randomly assigned to two different groups, which are referred to as Group A and Group B, of size $N_A$ and $N_B$, respectively. Students in Group A are assigned the exercises in Book A as homework during the course of a month, and students in Group B are assigned the exercises in Book B as homework during the same period of time. Students solve the exercises individually and are not allowed to interact with one another. The researcher is interested in establishing whether children who are assigned the exercises in one book perform better in a math examination at the end of the experiment than children assigned the exercises in the other book, but based on the available information, the researcher has no prior belief as to which book is more effective than the other. In this case, a two-sided hypothesis test is appropriate, because the alternative hypothesis should be left unspecified.

Students are given a math examination at the beginning and at the end of the experiment, and the change in test scores is recorded for each student. Assuming that the difference in test scores is approximately normally distributed with means $\mu_{\mathrm{A}}$ and $\mu_{\mathrm{B}}$ in Groups A and B, respectively, and equal variance, the mean differential effect of the two types of exercises can be analyzed using a two-sided difference-in-means test to determine whether $\mu_{\mathrm{A}} - \mu_{\mathrm{B}}$ is different from zero. Formally, the null and alternative hypotheses are formulated as follows:

$$H_0 : \mu_{\mathrm{A}} - \mu_{\mathrm{B}} = 0,$$

$$H_1 : \mu_{\mathrm{A}} - \mu_{\mathrm{B}} \neq 0.$$

The researcher chooses to test $H_0$ using the test statistic

$$W = \frac{\overline{X}_{\mathrm{A}} - \overline{X}_{\mathrm{B}}}{\sqrt{s^2\left(\frac{1}{N_{\mathrm{A}}} + \frac{1}{N_{\mathrm{B}}}\right)}},$$

where

$$s^2 = \left(\frac{1}{N_{\mathrm{A}} + N_{\mathrm{B}} - 2}\right) \left[\sum_{i=1}^{N_{\mathrm{A}}} \left(X_{i\mathrm{A}} - \overline{X}_{\mathrm{A}}\right)^2 + \sum_{i=1}^{N_{\mathrm{B}}} \left(X_{i\mathrm{B}} - \overline{X}_{\mathrm{B}}\right)^2\right]$$

where $\overline{X}_{\mathrm{A}}$ and $\overline{X}_{\mathrm{B}}$ are the sample means of the change in test scores in Groups A and B, respectively, and $X_{i\mathrm{A}}$ and $X_{i\mathrm{B}}$ are the changes in test scores for student $i$ in each group. $W$ is the $t$ statistic for the difference in means when variances are unknown but equal and has a $t$-distribution under $H_0$ with $N_{\mathrm{A}} + N_{\mathrm{B}} - 2$ degrees of freedom. However, because the number of degrees of freedom in this example is large ($N_{\mathrm{A}} + N_{\mathrm{B}} - 2 = 133$), the distribution of $W$ can be approximated by a normal distribution.

Assuming there are 70 students in Group A and 65 students in Group B and that $\overline{X}_{\mathrm{A}} = 0.1286$, $\overline{X}_{\mathrm{B}} = 0.0461$, $\sum_{i=1}^{N_{\mathrm{A}}} \left(X_{i\mathrm{A}} - \overline{X}_{\mathrm{A}}\right)^2 = 7.8429$, $\sum_{i=1}^{N_{\mathrm{B}}} (X_{i\mathrm{B}} - \overline{X}_{\mathrm{B}})^2 = 2.8615$, the value of the test statistic is $W = 1.6866$. To decide whether to reject the null hypothesis that both types of exercises have the same effect at increasing the math skills of students as measured by the improvement in their test scores, the significance level of the test must be established. If the significance level is set at 5% and this mass is equally distributed on both tails, the rejection rule is

Reject $H_0$ if $|W(X)| > 1.96$

Do not reject $H_0$ if $|W(X)| \leq 1.96$

since 1.96 and –1.96 are, respectively, the 2.5% and 97.5% quantiles of the normal distribution. Given that $|W(X)| = 1.6866 < 1.96$, $H_0$ cannot be rejected and the researcher cannot reject the hypothesis that exercises in Book A are equally effective at improving math skills than exercises in Book B.

In this example, had the researcher performed a one-sided test with the alternative hypothesis that $\mu_A - \mu_B > 0$, the rejection rule would have been

Reject $H_0$ if $W(X) > 1.64$,

Do not reject $H_0$ if $W(X) \leq 1.64$,

and the null hypothesis would have been rejected in favor of the alternative hypothesis that the type of exercises in Book A are more effective at increasing the mathematical skills of students than the type of exercises in Book B. Thus, using a two-sided test, the null hypothesis cannot be rejected, even when a one-sided test (to the right) would have rejected the null hypothesis.

*Rocío Titiunik*

***See also*** *p* Value; Significance, Statistical; Significance Level, Concept of; Statistic; Test; Type I Error; Type II Error

**Further Readings**

Box, G. E. P., Hunter, J. S., & Hunter, W. G. (2005). *Statistics for experimenters. Design, innovation and discovery.* Hoboken, NJ: Wiley-Interscience.
Casella, G., & Berger, R. L. (2002). *Statistical inference.* Pacific Grove, CA: Duxbury Press.
Hogg, R. V., & Craig, A. T. (1995). *Introduction to mathematical statistics.* Upper Saddle River, NJ: Prentice Hall.
Lehmann, E. L. (1986). *Testing statistical hypotheses.* New York: Springer.
Lehmann, E. L. (1998). *Nonparametrics. Statistical methods based on ranks.* Upper Saddle River, NJ: Prentice Hall.
Mittelhammer, R. C. (1995). *Mathematical statistics for economics and business.* New York: Springer.
Stone, C. J. (1996). *A course in probability and statistics.* Belmont, CA: Duxbury Press.

# Type I Error

Hypothesis testing is one of the most widely used procedures in statistical decision making. However, it can result in several errors. This entry focuses on the Type I error, which occurs when a true hypothesis is wrongly rejected.

## Hypothesis Testing

Type I errors occur in statistics when hypothesis tests are used to make statistical decisions based on experimental data. In this decision process, a simple statement or null hypothesis is formulated on the true status of an unobserved phenomenon. At the same time, an alternative hypothesis is defined to reflect the opposite situation. For instance, a new drug is tested for its capacity to reduce high blood pressure. The null hypothesis states that the new drug does not change blood pressure. Any differences that are observed are explained by random variation in the measurement process. The alternative hypothesis can be formulated as "the drug reduces high blood pressure."

Once the appropriate hypotheses have been formulated, one would like to test whether the null can be rejected in favor of the alternative. At this point, an experiment can be designed, a relevant test statistic and its distribution under the null can be derived, and data can be collected. These elements are required in the process to arrive at one of the two following conclusions:

- The null hypothesis can be rejected in favor of the alternative because the observed pattern in the sample cannot be explained merely by random variation, or
- There is not enough evidence to reject the null in favor of the alternative because the observed pattern in the sample is most likely a result of chance.

When the statistical test favors a decision that is in agreement with reality, a correct decision has been made. However, this decision procedure can also result in several errors (see Table 1).

The null hypothesis can be wrongly rejected in favor of the alternative, even if the statement in the null is true. In this case, a Type I error has occurred. One can also fail to reject the null hypothesis when the alternative is true, which results in a Type II error. For instance, in the blood pressure example, a Type I error occurs if a statistically significant

**Table 1** Schematic Representation of Possible Combinations between Hypothesis and Test Outcome

| | | *Null hypothesis* | |
|---|---|---|---|
| | | *True* | *False* |
| *Test outcome* | Accepted | Correct | Type II error |
| | Rejected | Type I error | Correct |

effect of the drug is found when in reality the drug does not influence blood pressure.

## Why Type I Errors Occur

Hypothesis testing can be viewed as making an educated guess. Like with any other guess, it is possible to draw the wrong conclusion. The data that are used to validate the null hypothesis are drawn as a random sample from the study population. Just by chance, it is possible that this sample reflects a relationship that is not present in the population. Consider for instance four throws with a six-sided die. If the four throws result in four 1s, this could be interpreted as evidence that the die is loaded in favor of the outcome "1." However, even with an honest die, a combination of four 1s can occur once in every 1,296 ($6^4$) sequences of throws. If in this case the die would be marked as loaded, a Type I error has occurred. Unfortunately, one can never be certain whether a rejected hypothesis is a reflection of the true nature of the studied phenomenon or whether it is the result of an accidental relationship present in the sample.

Depending on the context, making a Type I error can have important consequences. In the early stage of the development of a new drug, a Type I error would imply that a significant effect is found when there is none. Although this error could be detected in a later stage when a more complicated trial has been implemented, this might come at a considerable economical expense for the company developing the drug. A Type I error is, therefore, often considered to be a serious error: If there is an effect detected, then one wants to be as sure as possible that the effect is not the result of random variation.

In practice, one will try to control for the proportion of errors that could be made. The hypothesis testing starts by specifying $\alpha$, the probability of making a Type I error. This probability, which is also referred to as the significance level, is chosen as low as possible. Conventionally, a significance level of 5% or less is considered small enough to ensure that the result did not occur by chance. However, the user of this procedure should still be aware that a conclusion based on this significance level will still be wrong one in 20 times. Reducing the probability of observing a Type I error can easily be achieved by reducing the significance level of the test. If a significance level of 1% rather than 5% is selected, then the conclusion will be wrong only once in 100 times. If a significance level is set equal to 0, then the null hypothesis will never be wrongly rejected. In fact, in this case the null hypothesis will never be rejected at all. However, choosing a smaller significance level will also reduce the power of the test: Rejecting the null hypothesis will become more difficult, even when the alternative is true, unless you can also increase the sample size. As a result, the probability of observing a Type II error will be inflated. Because the probabilities of observing a Type I and a Type II error are inversely related, it might not always be easy to find a balance between the two.

## What Is in a Null?

Observe that the classification of an error as a Type I or Type II depends on the definition of the null hypothesis. Consider, for instance, the blood pressure example, with reversed hypotheses: The null is formulated in terms of the experimental drug having an effect, and the alternative as the treatment not having an effect. In this case a Type I error is made when treatment is found not to have an effect when in fact it does affect blood pressure. This error was previously labeled as Type II.

The choice of the hypotheses mainly depends on the interest of the study. Recall that hypothesis testing can result in one of two responses: Either the null hypothesis is rejected in favor of the alternative or it is not. Note that failing to reject the null is not the same as claiming that the null is true. One fails to reject the null when there is not enough evidence against it. Therefore, in conventional statistical practice, what one wants to prove is formulated as the alternative hypothesis. Technically, one is testing for "guilt" rather than "innocence." Like in any judiciary system, a defendant (hypothesis) is considered innocent (true) until proven guilty (false) beyond a reasonable doubt.

## Cumulative Type I Error Rates

A situation where Type I errors can become an important problem when not properly accounted for is when several hypotheses are tested simultaneously. Even though the probability of observing a Type I error within one test might be set at 5%,

if several tests are performed simultaneously, then the probability of observing a Type I error in at least one of these tests will be higher than 5%. Indeed, the possibility of observing a Type I error occurs each time a statistical test is performed. The more tests are performed, the higher the probability that a false positive significant result is found in at least one of them, just by chance.

The impact of multiple testing on the overall Type I error rate can be determined as follows. Let $\alpha_t$ represent the Type I error rate of the individual test, and let $k$ refer to the number of independent comparisons. Then, the overall Type I error rate can be approximated by the correction $1 - (1 - \alpha_t)^k$. For example, consider a hypothesis test with significance level set at 5%. When this test is performed 5 times simultaneously, the overall Type I error rate (i.e., the probability of observing a Type I error in at least one of the 5 tests) is inflated up to 22%. When this test is performed 10 times, the overall Type I error rate is inflated up to 40%; when this test is performed 20 times, the overall Type I error rate is inflated up to 64%, and so on. The more comparisons are performed, the higher the probability of observing a false positive result in at least one of them.

Still, it is possible to protect an analysis against an inflated Type I error rate resulting from multiple comparison testing. The approach is based on the idea that if the Type I error rate of each comparison is reduced sufficiently, then the overall Type I error rate will also decrease. For instance, the Bonferroni procedure consists of resizing the $\alpha_t$ for each individual comparison by dividing it by the number of comparisons. This strategy guarantees that the overall Type I error rate will not exceed the prespecified significance level. However, as mentioned before, reducing the probability of observing a Type I error for each individual test will make it more difficult to find significant results and will increase the probability of observing a Type II error. A bigger sample size might be required to increase the power of the test.

## Yes or No

Type I errors occur in hypothesis testing as part of a testing procedure that results in a yes-or-no answer. One either rejects the null hypothesis or one concludes that there is not enough evidence against it. This conclusion is often based on the $p$ value associated with the estimated test statistic. $p$ values represent the probability of observing data at least as extreme as the data that were actually observed, given that the null hypothesis is true. Therefore, a $p$ value can be interpreted as an indicator of the amount of evidence against the null: The smaller the $p$ value, the stronger the evidence. In hypothesis testing, the $p$ value is often compared with the prespecified significance level. This significance level is used as a cutoff point, where all $p$ values below result in a rejection of the null, and all values above are interpreted as compatible with the null. The main disadvantage of this procedure can be illustrated as follows. If a $p$ value of .06 is obtained from a test with a 5% significance level, then the null hypothesis will not be rejected, even though the evidence against it is strong. Clearly, reporting this $p$ value rather than the rejection of the null on the 5% significance level would be much more informative.

## Type I Errors in Real Life

Type I errors also frequently occur beyond the boundaries of statistical testing. Similar errors are observed in diagnostic tests. For instance, a pregnancy test could produce a positive result even in the absence of pregnancy. Note that with this type of diagnostic screening tests, Type II errors are often considered to be more serious than Type I errors. When tested positive, the test subject can be redirected to a more elaborate evaluation in which the initial positive result can be confirmed or not. In case a Type I error or a false positive result occurs at the time of screening, it can still be corrected at a later stage. However, in the opposite situation where the screening fails to detect the positive result, the test subject might miss an important opportunity to receive appropriate care.

A false positive can also occur when a legitimate e-mail is mistakenly labeled as spam by a filter or when a smoke alarm goes off without a fire. An innocent person could be found guilty of a crime that he or she did not commit. In fact, in this case a Type I error would even be twice as bad. Not only does an innocent person have to go to jail, but the guilty party goes free as well.

*Saskia Litière*

*See also* Bonferroni Procedure; Multiple Comparison Tests; Power; *p* Value; Sample Size Planning; Significance Level, Concept of; Type II Error

**Further Readings**

Freedman, D., Pisani, R., & Purves, R. (2007). *Statistics* (IS ed.). New York: W. W. Norton.

Keirse, M. J., & Hanssens, M. (2000). Control of error in randomized clinical trials. *European Journal of Obstetrics & Gynecology, 92,* 67–74.

# Type II Error

Hypothesis testing is one of the most widely used quantitative methods in decision making. It answers a research question in terms of statistical (non-) significance of a null hypothesis. The procedure of hypothesis testing can result in several errors. This entry focuses on Type II errors, which occur when a false hypothesis is not rejected. A short introduction to hypothesis testing is provided, followed by an overview of factors influencing the occurrence of Type II errors and examples of Type II errors beyond the boundaries of statistics.

## Hypothesis Testing

Developed by Jerzy Neyman and Egon S. Pearson, hypothesis testing has become one of the most widely used quantitative methodologies in almost all areas dealing with experiments and data. In this decision process, a simple statement or null hypothesis is formulated on the true status of an unobserved phenomenon. At the same time, an alternative hypothesis is defined to reflect the opposite situation. The outcome of the hypothesis testing can be one of the two following results:

- The null hypothesis can be rejected in favor of the alternative because the observed pattern in the sampled data cannot be explained merely by random variation.
- There is not enough evidence to reject the null in favor of the alternative because the observed pattern in the sampled data is most likely a result of chance.

Note that failing to reject the null is not the same as claiming that the null is true. In principle, one is testing for "guilt" rather than "innocence": Like in any judiciary system, a defendant (hypothesis) is considered innocent (true) until proven guilty (false) beyond any reasonable doubt. The alternative hypothesis will, therefore, always reflect the property that one would like to prove. Consider, for instance, a clinical trial comparing an experimental drug with a standard treatment. The purpose of the trial is to detect a difference in the effect of both drugs. The null hypothesis will state that the experimental drug is, on average, not better than the conventional drug. Any differences that are observed can be explained by random variation in the measurement process. The alternative hypothesis can be formulated in several ways. For instance, a two-sided alternative could state that the experimental drug has, on average, a different effect. If one expects to find an improvement, then a one-sided alternative like "the new drug is better, on average, than the conventional drug" can be considered.

Once the appropriate hypotheses have been formulated, one would like to test whether the null can be rejected in favor of the alternative. At this point, an experiment can be designed, a relevant test statistic and its distribution under the null can be derived, and data can be collected. When the test favors a decision that is in agreement with reality, a correct decision has been made. However, this decision process can also result in several errors (see Table 1).

On the one hand, a Type I error occurs when the null hypothesis is wrongly rejected in favor of the alternative. On the other hand, a Type II error is made when a false null hypothesis is not rejected. In the clinical trial example, a Type II

**Table 1** Schematic Representation of Possible Combinations between Hypothesis and Test Outcome

| | | *Null hypothesis* | |
|---|---|---|---|
| | | *True* | *False* |
| *Test outcome* | Accepted | Correct | Type II error |
| | Rejected | Type I error | Correct |

error occurs if there is not enough evidence against the null, when in reality the drug has an effect different from that of the conventional treatment.

## Factors Influencing Occurrence

In practice, one tries to control for the proportion of errors that could occur. Controlling for the probability $\alpha$ of making a Type I error can be done relatively easily by specifying it at the start of the study. This probability, which is also referred to as the significance level, is chosen as low as possible. Controlling for the probability $\beta$ of making a Type II error is less straightforward as $\beta$ depends on the prespecified significance level, the sample size, and the effect size one wishes to detect.

In clinical trials, the probability of observing a Type I error is often prespecified at 5%. It was a convenient choice in times when researchers had to rely on Ronald A. Fisher's tables of critical values based on 5% and 1% significance levels. Through the years, statisticians have continued to use 5%, almost like a scientific dogma, even in the current age of computers and software that greatly facilitate statistical computing. This level is considered small enough to ensure that the result did not occur by chance. Still, one should be aware that a conclusion based on a 5% significance level will still be wrong 1 in 20 times. Reducing the probability of observing a Type I error can be achieved relatively easily. The investigator can choose to lessen the significance level associated with the test. If a significance level of 1% rather than 5% is selected, then the conclusion will be wrong only 1 in 100 times. Choosing a smaller significance level does, however, have the adverse effect of complicating the rejection of the null hypothesis, even when the alternative is true. As a result, reducing the Type I error rate will inflate the probability of observing a Type II error.

Still, for any specified level of significance, the probability of observing a Type II error can be reduced by increasing the sample size. Indeed, adding information to the test implies adding evidence to the decision process. This strategy will result in a more informed conclusion. Several reexaminations of "negative" (or nonsignificant) trials have shown that in many studies, Type II errors occurred because of undersampling. Too few subjects were included in the study to guarantee a sufficient power to detect a clinically relevant effect. Note that the power of a test corresponds to $1-\beta$ (i.e., one minus the probability of observing a Type II error). When the power of the test is low, then a Type II error is more likely, and vice versa.

Nowadays, sample size calculations based on prespecified levels for the Type I and Type II errors have become common practice in clinical trials. Most investigators will accept a $\beta$ level of around 0.2 or, equivalently, a power of 80%. That means that in 8 of 10 trials, a statistically significant effect will be observed if the effect is in reality as large as the one that was presupposed. This also implies that in 2 of 10 trials, the null hypothesis will not be rejected in favor of the correct alternative, and it will therefore result in a Type II error. Selecting a Type II error rate equal to 0.1 or, equivalently, increasing the power of the test to 90% might require a substantial increase in the number of participants. This might not always be feasible from an economical point of view. Furthermore, if the phenomenon one wishes to study is infrequent, like a rare disease, it might not be practical either.

The probability of observing a Type II error also depends on the effect size that needs to be detected. Indeed, the smaller the effect, the lower the power of the test and, therefore, also the higher the probability of making a Type II error. As a result, a larger sample is required to detect a smaller effect.

Note that even if the null hypothesis is true, it will be unlikely that the observed effect size will be exactly zero. Random variation in the measurement process can cause some small deviations. If, in this case, the sample size of the study is chosen large enough, it will be possible to detect even this random variation. However, this statistically significant result would clearly not be scientifically relevant. Therefore, it should be decided at the start of a study what is the minimum effect size one wishes to detect. For instance, in the clinical trial example, detecting a 20% difference between the experimental drug and the conventional treatment might be more clinically relevant than detecting a difference of only 2%.

Furthermore, if the effect size is expected to be in a certain direction, then the alternative can also be defined as a one-sided statement. For instance, the new drug could be expected to perform at least as well as the standard treatment. Given that the

power to detect a one-sided alternative is higher than that of detecting a two-sided alternative, choosing one over the other can also have a positive impact on the probability of observing a Type II error.

Once a meaningful difference is agreed on, together with an acceptable Type I and Type II error rate, an appropriate sample size can be calculated to perform a successful trial. The hypothesis testing then serves as a validation tool that the significant findings did not occur by chance.

The decision of whether a finding is statistically significant based on the relatively arbitrary 5% level has been the topic of many passionate discussions among statisticians since the 1900s. Type I and Type II errors occur only because of the way hypothesis testing was developed. Neyman and Pearson motivated their approach by stating that, without knowing which hypothesis is true, hypothesis testing provides a rule to guide the decision process, hoping that in the long run, one will not be too often wrong. It does, however, not provide any information on whether the user will be wrong in one specific study. An alternative approach that does provide information on the current study is based on Fisher's $p$ value. $p$ values can be interpreted as the probability of observing data at least as extreme as the data that were actually observed, given that the null hypothesis is true. Therefore, a $p$ value is an indicator of the amount of evidence against the null: The smaller the $p$ value, the stronger the evidence. Neyman and Pearson actually used Fisher's idea in hypothesis testing, by comparing it with the Type I error rate: All $p$ values below $\alpha$ result in the rejection of the null hypothesis. Incidentally, this can lead to situations in which one would ignore a potentially clinically relevant result because a $p$ value of .06 was observed. If this $p$ value were to be reported rather than a conclusion reflecting the result of the hypothesis testing, the reader could get a better impression of the amount of evidence against the null hypothesis. Given his knowledge about the field and previous studies, it could help him or her to make a more informed decision.

### Beyond the Boundaries of Statistics

As the comparison of hypothesis testing with a judiciary system already illustrated, Type II errors are also frequently observed beyond the boundaries of statistics. A Type II error occurs when junk mail is not blocked by a spam filter or when a guilty person is not condemned by a lack of evidence. In some situations, Type II errors can even have serious consequences. For instance, a virus would not be detected by an antivirus system or a smoke alarm would not warn in case of fire. A pregnancy test can fail to detect pregnancy. With this type of diagnostic screening tests, Type II errors are often considered to be more serious than Type I errors. When testing positive for pregnancy or a disease like tuberculoses, Hepatitis A/B, or cancer to name but a few, the test subject will be referred to a more elaborate evaluation in which the initial positive result can be confirmed or denied. In case a Type I error or a false positive result occurs at the time of screening, it can still be corrected at a later stage. However, in the opposite situation in which the initial screening fails to detect the positive result, the test subject might miss an important opportunity to receive appropriate care.

*Saskia Litière*

***See also*** Power; $p$ Value; Sample Size Planning; Significance Level, Concept of; Type I Error

**Further Readings**

Freedman, D., Pisani, R., & Purves, R. (2007). *Statistics* (IS ed.). New York: W. W. Norton.

Keirse, M. J., & Hanssens, M. (2000). Control of error in randomized clinical trials. *European Journal of Obstetrics & Gynecology, 92,* 67–74.

Lehman, E. L. (1993). The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? *Journal of the American Statistical Association, 88,* 1242–1249.

## Type III Error

Type III error has been defined in two primary ways. First, researchers have described a Type III error as occurring when a research study provides the right answer but for the wrong question or research hypothesis. For example, in public health research, when a research hypothesis predicts risk

differences between groups or across time periods, the study design requires adequate characterization of all relevant groups and time periods. If discrepancies are found among the research hypothesis, time periods, and the methods used to test the hypothesis, these discrepancies can lead to what seems to be answers to a specific question, when in fact the results support a different question. Second, other researchers have stated that in statistical tests involving directional decisions, a Type III error can occur when one accepts a specific directional alternative when in fact another alternative is true. In this case, the Type III error has a conditional probability, gamma, which is simply the probability of getting the direction of the effect wrong. The entry provides some examples of Type III errors and recommendations for avoiding Type III errors.

## Examples

### *Stroke*

As noted in the first definition, Type III errors can occur when one examines the causes of variation in risk for a disease between individuals, but it does not consider that the causes of interindividual risk might differ from the causes of disease rates over time, and from causes of rate differences between populations. For example, when determining the etiology of stroke, the sources of variation of risk might include differences between individuals in genetic vulnerability, current medical conditions, temporal changes in poverty, and access to health care, as well as population and societal differences in what type of health care system is available. Furthermore, each of these factors is distinct from the proximal cause of stroke, namely the deprivation of oxygen to the brain. Certainly, each of the factors discussed previously contributes to the risk for stroke. A Type III error can occur when a study is designed to ascertain the inter-individual differences in risk for stroke within a given population, but it neglects to look for causes of stroke that are constant within the population being studied. A Type III error can also occur when neglecting risk factors for stroke related to variation between populations and across time. Furthermore, a Type III error might occur when one determines the risk for stroke in more than one population and finds interindividual differences in risk of stroke that produce variation of stroke outcomes within one population, but it fails to realize that the causes that produce interindividual differences in risk for stroke in one population might be different from the factors that produced differences in average risk for stroke between two or more populations.

### *Interactive Effects From Rehabilitation After Stroke*

In a large, multisite, randomized controlled trial using leisure therapy and conventional occupational therapy administered to patients who have had a stroke, Chris Parker and colleagues failed to demonstrate a benefit for patients in terms of improved mood, leisure activity, or independence in activities of daily living measured at 6 and 12 months. Based on the second definition of Type III error, one possible explanation of the failure to demonstrate an effect in this study might be the failure to consider interactive effects of the interventions, which might have produced a Type III error in interpretation of the results.

## Recommendations

To avoid a Type III error when interpreting the findings from adequately powered studies that use multiple interventions, one must be careful not to conclude that a negative finding indicates that all components of the intervention were ineffective. Also, one should reassess whether the study design adequately measures potential interactive effects that occur with the intervention. Furthermore, a negative finding should not cause one to overreact and ignore one's findings or to ignore plausible alterative hypotheses.

*Michael A. Dawes*

***See also*** Power; *p* Value; Sample Size Planning; Significance Level, Concept of; Type I Error; Type II Error

### Further Readings

Leventhal, L., & Huynh, C.-L. (1996). Directional decisions for two-tailed tests: Power, error rates, and sample size. *Psychological Methods, 3,* 278–292.

Parker, C. J., Gladman, J. R. F., Drummond, A. E. R., Dewey, M. E., Lincoln, N. B., Barer, D., et al. (2001). A multicentre randomized controlled trial of leisure therapy and conventional occupational therapy after stroke. *Clinical Rehabilitation, 15,* 42–52.

Schwartz, S., & Carpenter, K. M. (1999). The right answer for the wrong question: Consequences of Type III error for public health research. *American Journal of Public Health, 89,* 1175–1180.

Wade, D. T. (2001). Research into the black box of rehabilitation: The risks of a Type III error. *Clinical Rehabilitation, 15,* 1–4.

# U

## Unbiased Estimator

One of the important objectives of scientific studies is to estimate quantities of interest from a population of subjects. These quantities are called parameters, and the actual nature of the parameter varies from population to population and from study to study. For example, in a study designed to estimate the proportion of U.S. citizens who approve the President's economic stimulus plan, the parameter of interest is a proportion. In other cases, parameters may be the population mean, variance, median, and so on. In cases where the population values are represented by a parametric model, the parameters represent the whole population instead of a particular characteristic of it.

Parameters are estimated based on the sample values. Estimators are functions of sample observations used to estimate the parameter. It would be expected that a good estimator should result in an estimate that is close to the true value of the parameter. Because in practice the parameters are unknown, it is not possible to compare the estimate with the true value. Unbiasedness serves as a measure of closeness between an estimator and the parameter.

Suppose the population parameter $\theta$ is to be estimated based on the sample observations $X_1, X_2, \ldots, X_n$. An estimator $\hat{\theta} = h(X_1, X_2, \ldots, X_n)$ is an unbiased estimator of $\theta$ if the mean or expectation of $\hat{\theta}$ over all possible samples from the population results in $\theta$. In statistical terms, $\hat{\theta} = h(X_1, X_2, \ldots, X_n)$ is an unbiased estimator of $\theta$ if and only if

$$\mathrm{E}(\hat{\theta}) = \theta$$

for all $\theta$, where $\mathrm{E}(\hat{\theta})$ represents the expected value of $\hat{\theta}$. That is, if in repeated sampling from the population, the average value of the estimates equals the true parameter value, then the estimator is unbiased.

As an example, suppose the population mean $\mu$ is to be estimated based on a random sample $X_1, X_2, \ldots, X_n$ from the population, then the sample mean $\overline{X} = (X_1 + X_2 + \cdots + X_n)/n$ is an unbiased estimator of the population mean. Statistically,

$$\begin{aligned} E(\overline{X}) &= \frac{E(X_1 + X_2 + \cdots + X_n)}{n} \\ &= \frac{[E(X_1) + E(X_2) + \cdots + E(X_n)]}{n}, \\ &= \frac{\mu + \mu + \cdots + \mu}{n} = \mu, \end{aligned}$$

for all $\mu$. To see this, a pathological example will be used. Suppose the population consists of only five subjects and the measurements on these five subjects are listed as follows:

$$3, 5, 3, 2, \text{and } 2,$$

resulting in population mean

$$\mu = (3 + 5 + 3 + 2 + 2)/5 = 3.$$

**Table 1** All Possible Samples and Corresponding Sample Means From the Population (3, 5, 3, 2, and 2)

| *Sample* | *Sample Mean* |
|---|---|
| (3; 5) | 4.0 |
| (3; 3) | 3.0 |
| (3; 2) | 2.5 |
| (3; 2) | 2.5 |
| (5; 3) | 4.0 |
| (5; 2) | 3.5 |
| (5; 2) | 3.5 |
| (3; 2) | 2.5 |
| (3; 2) | 2.5 |
| (2; 2) | 2.0 |

Now consider estimating $\mu$ based on without-replacement-samples, that is, samples containing distinct members of the population, of size 2 from this population. Table 1 shows possible samples and the corresponding sample means. Note that the average of the sample means is $(4.0 + 3.0 + 2.5 + 2.5 + 4.0 + 3.5 + 3.5 + 2.5 + 2.5 + 2.0)/10 = 30/10 = 3 = \mu$.

Similarly, a sample proportion is an unbiased estimator of the population proportion. In general, the sample median is not an unbiased estimator of the population median. Sample variance, defined by the formula

$$S^2 = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n-1}$$

is an unbiased estimator of the population variance $\sigma^2$.

When the population is infinite or large, it is not possible to investigate each and every member of the population to determine the population characteristics, referred to as parameters. An unbiased estimator of a population parameter closely approximates the parameter. Without evaluating the whole population, the population parameter can be computed with accuracy based on the unbiased estimator from a sample drawn from the population. This is because in repeated sampling, the unbiased estimator results in an average value that is equal to the parameter itself.

*Abdus S. Wahed*

*See also* Biased Estimator; Estimation; Mean; Median; Parameters; Population

**Further Readings**

Rice, J. A. (1994). *Mathematical statistics and data analysis*. Belmont, CA: Duxbury.

# Unit of Analysis

When engaging in research, it is essential to have a clear understanding from the outset what the study is trying to accomplish. That is, what the specific research question(s) is(are), and what data are necessary to collect to answer the research question(s). Otherwise, confusion typically ensues. One of the most fundamental considerations in conducting research is to determine what the primary unit that will be the subject of statistical analysis is, or should be. This is called the unit of analysis. Often, it is dictated by the data that are collected, rather than by a sound theoretical justification. Therefore, it is necessary to be clear as to what the unit of analysis should be before data collection. For example, in educational research, the unit of analysis could be people or schools. The unit of analysis may be the same or different from the unit of generalization, the unit of sampling, or the unit of measurement.

To understand the unit of analysis, it is essential to understand these related terms. An example may help clarify the distinction between the terms. Consider the case where a researcher seeks to investigate the effect of a curriculum on student achievement. Data can be collected at a variety of levels: the student level, the classroom level, the school level, the district level, the state level, and so on. Here, the example is restricted to the first three levels: the student level, the classroom level, and the school level. Similarly, the researcher can sample schools, classrooms, or students to collect the data. The choice of which unit to choose for sampling and data collection depends, in part, on the unit of generalization. To understand the unit of generalization, it is useful to understand what it would mean to generalize at the three different levels. If the study seeks to generalize at the school level, that would mean that the researcher is

interested in how implementing the new curriculum effects the achievement of students at the school level. The interest is not on individual achievement but on the average achievement for the school. Similarly, the researcher may wish to generalize to the classroom level, at which the results would be reported for each classroom. Last, the goal could be to determine the effect of the instruction at the individual level, and hence, the results of the individual would be paramount. Therefore, if the goal were to generalize at the school level, for instance, it would be wise to do the sampling at the school level, to be sure that there are enough schools on which valid inferences can be made, and that the sample of schools is representative. If the goal is at the individual student level, in contrast, then it may be possible to gather enough students on which to base valid inferences, and also that is representative, without having enough schools for a valid inference to be made. Therefore, prior to sampling and data collection, the unit of generalization should be determined.

It is typical that the unit of sampling and the unit of generalization are the same, but they may be different from the unit of measurement. Suppose in the example that the unit of generalization is to be schools. Therefore, the results will be school-level achievement results, and the sampling will likely occur at the school level. However, the unit of measurement will likely take place at the student level. To determine the school-level achievement, it is necessary to collect data on student achievement and possibly then to aggregate that student-level data into school-level data, perhaps as the mean of the student-level results. It may be that other data are collected as well, not only achievement data on students, and as such data collection could take place at any and all of the three levels: at the student level, classroom level, and school level. At the student level, data may be collected on not only the achievement of the student but also on other variables such as the gender of the student, the ethnicity of the student, and so on. At the classroom level, data regarding the number of students in a classroom, the number of years of experience of the teacher, and the gender of the teacher may be collected. And at the school level, data regarding the total enrollment of the school, the demographic makeup of the school, the student-to-teacher ratio, and so on, can be collected. All these data can be used to assess the school-level performance. Therefore, although the unit of sampling was the school, the unit of measurement may be any one of the three levels, or some combination thereof. In this case, the unit of measurement is different from the unit of generalization and from the unit of sampling.

Typically, the unit of analysis is the same as the unit of generalization, although there are cases where this is not true. In the example, data are collected at the individual student level (and perhaps at the classroom level and the school level as well). These data are then aggregated to school-level data. In the simplest case, data could be collected on student achievement alone, and these results may be averaged for the school, resulting in a mean achievement for the school. This variable is a school-level variable and is what is used in the analysis for the study. It is worth noting that when the unit of analysis is the school, different data are required than would be necessary than if the unit of analysis were the student. In the most simplistic case, it is instructive to think about the mean of the achievement scores of the students. Because the mean is more stable than the individual score, valid inferences can result from analyzing a smaller number of means than would be advised when drawing inferences from individual observations or scores. So although it may be valid to draw inferences from 30 means, each based on 200 students, it would not usually be reasonable to use 30 individual observations to make valid inferences at the individual level.

When choosing the appropriate unit of analysis, the researcher should also be aware of the ecological fallacy, which states that the conclusion(s) drawn at the group (in this case, school) level may not pertain to the individual (i.e., student), and conversely, that the conclusion(s) drawn based on the analysis of individual (i.e., student) level may not be accurate at the group (i.e., school) level. Therefore, when choosing the unit of analysis, it is essential to consider the unit of generalization, as the conclusions and inferences drawn as a result of the analysis may be accurate only at the level of the unit of analysis. Thus, if it is desired to draw conclusions about achievement at the school level, the unit of analysis, and generalization should be at the school level, not at the student level, regardless of the unit of measurement.

Thus far, the discussion of the unit of analysis has been relatively straightforward; basically, the unit of analysis is chosen as the unit of generalization, which is typically the unit of sampling as well. However, the unit of measurement may not be the same as the unit of sampling, generalization, and analysis. One important factor has been neglected in this discussion, namely that of the independence of observations. In many statistical analyses, an assumption of the independence of observations is made, and when data are nested, such as students within classrooms within schools, observations at one level may not be independent, whereas observations at another level may be. For instance, there may not be an independence of observations of students within the same classroom because the students share the common experience of that classroom. Conceptually, it can be made obvious by considering the results of students within a classroom relative to those in different classrooms. If students within a classroom perform more similarly to each other than they do to students in a different classroom, then there is a classroom-level effect in play and the independence of the observations at the classroom level is suspect. If, however, it can be assumed that the classrooms are independent, then classroom-level data should be used in the analysis. If being in the same school causes a dependence of measurement between classrooms, it would be necessary to aggregate classrooms into a school-level variable. This could continue until the observations for each unit are considered independent. Thus, when observations are independent at the school level, the unit of analysis should be the school level. Because of the ecological fallacy, the unit of generalization must also be at the school level. If it is desired to have the unit of analysis and generalization to be at the student level, then students who are from independent units can be sampled. That is, if there is dependency at the classroom level, but not at the school level, then students can be sampled from different schools for the analysis.

Methods that do not require the independence of units, such as hierarchical models, render the unit of analysis a nonissue. In this case, all units are used as units of analysis. In the preceding example, the unit of analysis at the first level is the student, at the second level it is the classroom, and at the third level it is the school. In this case, the dependence among the units is modeled (and estimated). Conclusions can then be drawn at any level using the appropriate results.

*Lisa A. Keller*

***See also*** Mean; Observations; Sampling; Variable

**Further Readings**

Kenny, D. A. (1996). The design and analysis of social-interaction research. *Annual Review of Psychology, 47,* 59–86.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review, 15,* 351–357.

## U-Shaped Curve

The U-shaped curve usually refers to the nonlinear relationship between two variables, in particular, a dependent and an independent variable. Because many analytic methods assume an underlying linear relationship, systematic deviation from linearity can lead to bias in estimation. Meaningful U-shaped relationships can be found in epidemiology (e.g., between risk factor and disease outcome or mortality), psychology (often age-related developments, such as delinquency or marital happiness), and economics (e.g., short-run cost curves between the variate cost and quantity).

In medicine, U-shaped risk curves have been found for risk factors such as cholesterol level, diastolic blood pressure, work stress, and alcohol use. Of these factors, the alleged U-shape relationship between alcohol use and disease risk has been the most controversial. By the 1920s, a U.S. study by Raymond Pearl already showed a depressed longevity for abstainers. At that time, with alcohol prohibition in effect, this was not a politically correct message. Many years later, better controlled cohort studies looking into what Alvan R. Feinstein in his *Science* article called the "menace of daily life" have also reported lowest risk estimates for light or moderate drinkers of alcoholic beverages. Heavier drinkers are at highest risk, as could be expected. However, abstainers or nondrinkers in general also are found to have a higher

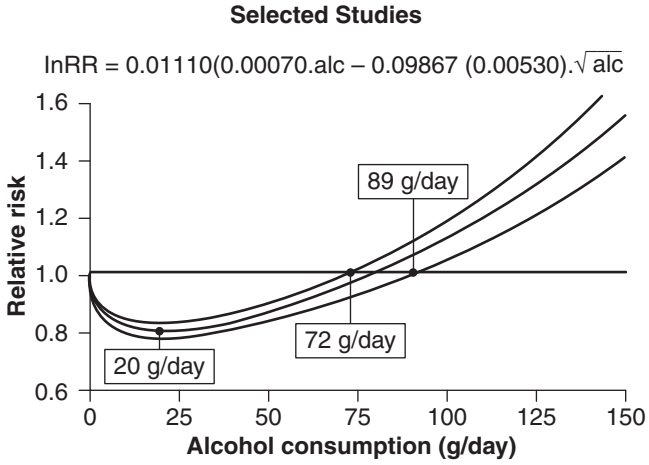



**Figure 1** Example of U- or J-Shaped Curve Between Alcohol Intake and Risk for Coronary Heart Disease

*Source:* Corrao, G., Rubbiati, L., Bagnardi, V., Zambon, A., & Poikolainen, K. (2000). Alcohol and coronary heart disease: A meta-analysis. *Addiction, 95,* 1505–1523.

risk for several negative health outcomes. This effect has been observed for overall mortality and specific categories such as cardiovascular diseases. For the latter, some studies report a J-shape rather than a U-shape, with little increased risk at higher consumption levels. Generally, the risk is estimated to be approximately 20% higher for abstainers, as shown in Figure 1 by Giovanni Corrao and colleagues.

During the last 20 years, as results from more and more cohort studies have been accumulating, the J-shaped risk curve has been considered to be the aggregate result of several biological processes underlying the most prevalent of pathologies in the Western world, coronary heart disease. For some diseases or bodily processes, any alcohol has an outright negative effect. Alcohol has been found to raise blood pressure even in small amounts, which in turn is a risk factor for cardiovascular disease. However, alcohol has a proven negative effect on the formation of thrombi or blood clots, which in itself is considered a risk factor for ischemic diseases (heart attack, brain infarctions). A major third process is the positive effect of alcohol use on the high-density cholesterol (HDL) level in the blood, which is considered to be a protective factor in the genesis of arterial plaques, eventually obstructing blood flow to vital tissues of heart or brain. Across the years, several other biological processes and genetic vulnerability factors have been suggested as potential candidates for the explanation of the lower risk for moderate drinkers of alcohol. The message of a potential beneficial health effect of alcohol use has caused considerable debate, as alcohol use at higher intake levels may be considered a serious health hazard. The detrimental effects of alcohol are less disputed, with monotonically increasing risk for outcomes such as injuries, liver functions, liver cirrhosis, and certain forms of cancer (e.g., breast cancer).

Next to the biological explanations of the J-shape, suggesting several direct (e.g., clotting) or indirect (high density cholesterol level) biological effects, some have provided alternative explanations, stemming from methodological flaws or specific design features. For example, with evidence from the British Regional Heart Study (BRHS), A. Gerald Shaper suggested that the lower effect could be the result of a mixture of nondrinkers with heterogeneous risk profiles. At the time, not all studies made a difference among lifetime abstainers, teetotalers, and ex-drinkers. From the BRHS cohort, it was obvious that the ex-drinkers were at the highest risk and that the risk of British teetotalers could not be distinguished from that of light drinkers. However, heterogeneity of risk in the nondrinker category cannot explain the U-shape at the lower end of the drinking scale.

Another explanation for the U- or J-shape at the lower end is the selection of high-risk individuals in the abstainer category, leading to what in epidemiology has been termed by Feinstein as *susceptibility bias*. Susceptibility bias would occur when persons more vulnerable to the disease would refrain from engaging in a drinking career. There is some evidence from a Dutch cross-sectional population study of a selection of young people with poorer health into the category of teetotalers, when health complaints among teetotalers increased with age as number of teetotalers in the population decreased with age. Solid empirical evidence for this phenomenon is difficult as it requires a follow-up of long duration of adolescent cohorts. In one such rare study of Bostonian teens, George

Eman Vaillant indeed presented evidence for selection leading to an overrepresentation of people with a higher disease burden among abstainers. A variant of this is termed *protopathic bias,* when people with preexisting disease or with a high-risk profile are overrepresented among the nondrinkers. Empirical evidence for these phenomena remains inconclusive, however.

Three other design features have been brought forward as potentially creating a bias in estimates of risk at the lower end of the alcohol consumption curve. In all epidemiologic studies, alcohol use (risk factor) is measured close to the endpoint (disease outcome) in cohorts that usually have passed middle age because incidence and mortality in younger age groups is quite low. Apart from a restricted time window of observation, this also means that risk assessments refer to recent alcohol intake at middle age, thereby disregarding the effects of drinking in the (average) first 20 to 30 years of life. Both aspects could produce a bias because changes in drinking status as a result of bad health would go unnoticed. Proof of such a bias was found in a New Zealand case-control study by Stijn Wouters and colleagues who found that cases were significantly more likely than controls to report recent abstention from drinking because they felt unwell. The U- or J-shaped association between recent alcohol consumption and acute coronary heart disease seemed to be largely caused by the confounding effect of preclinical, usually not measured, symptoms on drinking. Third, it has been reported in recent reviews that the lower end of the J-shape becomes less pronounced as quality of the design increases.

Apart from the caveats in assessing and interpreting the J- or U-shaped curve, estimation of the exact shape of the curve might present problems. Els J. T. Goetghebeur and Stuart J. Pocock have warned against oversimplification with potential bias in estimating the nadir (lowest turning point) and upward slope on particularly the left side of a U-shape. The often sparse numbers of observations available make exact fitting difficult, and categorization leads to loss of information. Several approaches to (parametric) fitting are possible. The authors suggest sequential quadratic and linear tests for the downward trend from the left of the curve.

*Paul Lemmens*

***See also*** Bias; Cohort Design; Confounding; Odds Ratio; Survival Analysis

**Further Readings**

Corrao, G., Rubbiati, L., Bagnardi, V., Zambon, A., & Poikolainen, K. (2000). Alcohol and coronary heart disease: A meta-analysis. *Addiction, 95,* 1505–1523.

Feinstein, A. R., & Horwitz, R. I. (1981). An algebraic analysis of biases due to exclusion, susceptibility, and protopathic prescription in case-control research. *Journal of Chronic Diseases, 34,* 393–403.

Goetghebeur, E. J. T., & Pocock, S. J. (1995). Detection and estimation of J-shaped risk-response relationships. *Journal of the Royal Statistical Society Series A, 158,* 107–121.

# V

## "Validity"

Validity is defined by Samuel Messick as an integrated, evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment. This definition opens Messick's chapter, "Validity," in the third edition of *Educational Measurement,* which is a benchmark publication in the field. Although Messick's conception of validity has generated some debate, this chapter has arguably provided the dominant scholarly representation of validity in educational measurement since its publication in 1989. Like the validity chapters in the preceding (1950, 1971) and subsequent (2006) editions of *Educational Measurement* (by Edward E. Cureton, Lee J. Cronbach, and Michael J. Kane, respectively), it is intended to provide guidance to those who develop and use tests or other assessments, as well as philosophical grounding for validity scholars. Key elements in Messick's representation of validity, which are elaborated in the following discussion, include its tie to scientific hypothesis testing and theory building, the need to address two general threats to validity with two general types of evidence (convergent and discriminant), and the importance of attending to several aspects of validity that illuminate sets of issues and sources of evidence. Issues surrounding Messick's conception of validity include his incorporation of consequences of testing as an aspect of validity and the extent to which his representation provides sufficient guidance for practitioners. This entry also discusses current issues regarding validity.

### Definitions and Scope of Application

Validation entails ascertaining the degree to which multiple lines of evidence are consonant with the intended inference (or "construct"), while establishing that alternative inferences are less well supported. This judgment takes into account the meaning of test scores as well as their relevance and utility for particular applied purposes, their value implications, and the social consequences of using them for applied decision making. Messick's definition of validity signals that validity is a property of inferences based on test scores, not tests themselves; that it is a matter of degree, not all or none; and that it evolves as new evidence is brought to bear.

He uses the term *test* to refer not just to tests as typically conceived but to any means of observing or documenting consistent responses, behaviors, or attributes; *scores* subsume qualitative and quantitative summaries of these consistencies and can refer to groups, situations, or objects, as well as persons. The term *construct* refers not just to unobservable qualities assumed to account for performance on a test, as originally conceived, but to whatever concept(s) the test is designed to measure. Thus, *construct, score meaning,* and *intended inference* or *interpretation* can be used interchangeably, as can

*construct validity* and *validity. Construct theory* encompasses (a) the definition of the construct (or "construct domain") that specifies the boundaries and facets of the construct as well as (b) the expectations (hypotheses) about responses processes, consistencies in item responses, relationships among variables, and so on.

## Key Elements in Messick's Conception of Validity

The fit (or lack of fit) between the construct theory and the available evidence informs the overall judgment about the validity of the intended inference. Thus, validation incorporates all the experimental, statistical, and philosophical means by which hypotheses and scientific theories are evaluated. Messick also acknowledges the need to balance the never-ending aspect of validation with the need to make the most responsible case to guide current use of the test.

Messick foregrounds two major threats to validity: *Construct underrepresentation* occurs when a test is too narrow in that it fails to capture important aspects of the construct; and *construct-irrelevant variance* occurs when a test is too broad in that it requires capabilities irrelevant or extraneous to the proposed construct. For instance, depending on the construct definition, a test of reading comprehension might be too narrow in that it focuses on short reading passages that are not typical of what students encounter in or out of school; or it might require irrelevant capabilities such as knowledge of specialized vocabulary. Such rival hypotheses to score meaning can be investigated by considering *convergent* and *discriminant* evidence. *Convergent* evidence indicates that test scores are related to other measures of the same construct and to other variables they should relate to as predicted by the construct theory; *discriminant* evidence indicates that test scores are not unduly related to measures of other constructs. For instance, scores on a multiple-choice test of reading comprehension might be expected to relate more closely (convergent evidence) to other measures of reading comprehension, perhaps using other texts or response formats. Conversely, test scores might be expected to relate less closely (discriminant evidence) to measures of the specific subject matter knowledge reflected in the reading passages on the test, indicating it is not primarily a measure of specialized knowledge.

Messick proposed six "aspects" or "components" of validity that cover a range of issues and sources of evidence to be considered in any test validation effort (foregrounded more distinctly in later publications). For each aspect, the goal is to ascertain the degree to which the evidence is consistent with the intended inference (anticipated by the construct theory) and less consistent with alternative inferences about score meaning. The *content* aspect considers the extent to which the content of the test is relevant to and representative of the construct definition. The *substantive* aspect considers the extent to which responses to the test are consistent with the construct definition; it includes evidence of the processes through which examinees respond to the test items and of consistencies among responses to different items. The *structural* aspect addresses the extent to which the scoring model is consistent with the construct theory; this includes attention to issues like how items are weighted or what dimensions are reflected in the scoring model. The *external* aspect addresses the extent to which the relationship between the test and other measures (of the same and different constructs) are consistent (appropriately high, low, or interactive) with what the construct theory would predict. The *generalizability* aspect addresses the extent to which score meaning is consistent across relevant contexts, including population groups, settings, time periods, and task domains. Finally, the *consequential* aspect traces the social consequences of interpreting and using the test scores in particular ways, scrutinizing not only the intended outcomes but also unintended side effects, considering their fit with what the construct theory would predict and any threats to validity they illuminate.

## Differing Conceptions of Validity

Messick's representation of validity reflects an evolution in the concept to which he and other scholars have contributed, albeit not without some controversy and continuing debate. Features of this evolution now reflected in the

widely endorsed *Standards for Educational and Psychological Testing* include (a) a move from different types of validity for different types of inferences (content, criterion, and construct validities) to a unitary concept of validity that combines multiples types of evidence into an overall judgment, (b) a more comprehensive set of aspects or categories of validity evidence (although Messick's "aspects" do not fully overlap the sources of validity evidence in the *Standards*), and (c) attention to the consequences of testing as an aspect of validity (which is arguably the most controversial change). Other issues surrounding Messick's work have included whether validity should refer to test-based *actions* as well as inferences, whether the meaning of scores can or should be evaluated separately from the uses to which they are put, and the extent to which his representation provides sufficient guidance for practitioners. These differing conceptions of validity and the issues that underlie them are addressed in the suggestions for further reading.

*Pamela A. Moss*

***See also*** Construct Validity; Content Validity; Criterion Validity

**Further Readings**

American Education Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education/ Praeger Publishers.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education/Macmillan.

Messick, S. (1996). Validity of performance assessments. In G. W. Phillips (Ed.), *Technical issues in large-scale performance assessment* (pp. 1–18). Washington, DC: U.S. Department of Education.

Moss, P. A., Girard, B., & Haniford, L. (2006). Validity in educational assessment. *Review of Research in Education, 30,* 109–162.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19,* 405–450.

# Validity of Measurement

In measurement, validity is traditionally defined as an argument in support of a construct, made using data. Although this definition might seem simple, its meaning changes as investigators adopt different scientific philosophies, descriptions of the entities to be measured, and norms for acceptable data. Uncertainty regarding what is to be measured has taught researchers about the fallibility of all forms of measurement. Rather than expect perfection, most researchers accept that new knowledge can be generated only if important ideas and events are assessed well enough to support meaningful conclusions. Assessment guidelines from three general approaches to measurement—labeled here as psychometric, developmental, and interpretive—reveal how investigators generate valid evidence and respond to the inevitable threats that might undermine even the strongest research design if left unchecked. Before outlining how validity is established, a brief representation of how validity norms differ across research paradigms illustrates some of the challenges associated with measuring anything well.

## The Strength of Validity

Perhaps one of the reasons it is difficult to understand measurement validity is the belief that science involves the discovery of truth. Truthful discoveries might be the ultimate aim of research; yet investigators examine a variety of truths and realities. Research in this information age is a process of fighting for language that will help individuals better understand one another and the universe in which they live. At best, new research can offer improvements in the accuracy with which phenomena are explained, and measurement is the step in the research process whereby investigators describe what they understand. Valid measurement arguments, in other words, can offer strong or weak descriptions of the concepts used to generate and defend new ideas.

With this in mind, it is helpful to remember that the relative strength of a validity argument is likely to differ depending on how well something is understood and on the qualities of the entities

being measured. Imagine two strong validity arguments that can be made for a well-understood construct. In the simplest case, just as a light switch can turn a bulb on and off, constructs might reflect a single dimension that can be assessed as present or absent, right or wrong. As occurs when electric frequencies vary, the underlying cause of such events might be complex, but indicators of these simple constructs are readily apparent and can be recorded accurately. More commonly in the social sciences, a construct might be well known but has complex properties. Strong validity arguments about such multidimensional constructs can be made accurately only if indicators of each dimension are included in the measurement instrument. Validity arguments about complex constructs include comparisons across dimensions as well as across the indicants of each dimension. Both simple and complex arguments are supported by evidence that all dimensions and indicants are recorded accurately and consistently across repeated uses of a measurement tool, and that the construct is captured fully.

Regardless of whether a construct has one or several dimensions, well-understood constructs are the easiest to measure because there are fewer opportunities for multiple interpretations of their existence. In mathematics, 2 + 2 will equal 4 in a base 10 system, and the accuracy with which individuals can answer such problems can be evaluated across repeated use of an addition test with relative ease. If the construct of addition were to be broadened to include other basic mathematical operations, more dimensions would be needed to represent fully the construct. Along with the inclusion of addition problems, for example, a measurement argument might include subtraction, multiplication, and division problems. A seemingly simple change in the definition of a construct invariably results in a dramatic revision of how validity is established.

Investigators who endorse a psychometric approach to measurement have designed the most specific notion of precision in how a construct is defined and assessed. These researchers regularly deliberate about how to define each construct as they endeavor to design instruments for use in a variety of situations and evaluate the stability and accuracy with which each construct is measured. Measurement in the absence of a well-defined construct is believed to yield a weak validity argument. Formulating strong validity arguments involves a continual revision of constructs and their measurements until all the assessment ambiguities are eliminated. The resulting arguments show evidence of accuracy, internal consistency, and stability in how a tool functions across persons and situations.

A second set of validity struggles was revealed by investigators who study change. Rather than endeavor to record precisely a well-defined construct, developmentalists expect variation in how a construct is defined over time, across people, and in different contexts. Focusing on processes, developmentalists study structural changes in how constructs are defined, temporal change in the presence or absence of such constructs, and individual and contextual variation in the expression of change. A psychometrician might argue that change can be measured if the definition of any meaningful construct is narrow enough to endure across studies, and developmentalists sometimes agree with this logic. That is, developmentalists sometimes recognize that change itself can be a construct. They measure such change by stringing together collections of specific constructs to support broader concepts. Nevertheless, developmentalists also assume that change can take many forms and that stringing together constructs would not facilitate a parsimonious assessment of change.

The quest for theoretical and empirical parsimony challenged developmentalists to invent new methods for describing predictable patterns of change. Evidence for such patterns might or might not depend on how the altered entities are defined. In some cases, a concept can become more complex over time or across contexts such that a unidimensional concept can gradually become multidimensional. In other cases, a multidimensional concept can become consolidated in a way that renders later measurements to be simpler than earlier measurements. It is also possible for the dimensions of an initial concept to become so altered when combined with other concepts that a new concept emerges. The definition of the concept might differ across studies, but strong validity arguments accommodate the dynamic processes that investigators are interested in understanding. These processes are revealed by anticipating the potential for change,

sometimes at the expense of fully articulating what changes.

This set of problems became most recently evident in research on brain functioning. For centuries, the measurement tools used to study how the brain works were designed with the idea that the brain itself did not grow much once an individual passed through early childhood. It was only once magnetic resonance imaging machines were invented and used to study situational functioning over time that researchers discovered cycles of new growth across the life span. In combination with thoughtful research designs, this instrument revealed periods during which the brain overproduces new cells and neurons and then prunes this growth as some cells are used and others atrophy. Such changes in brain functioning were not anticipated and emerged only as an artifact of attempts to map brain functioning in response to specific problems. Few investigators anticipated that the design of a new measurement tool could so profoundly change our understanding of human functioning; yet the instrument also has limits. All strong validity arguments openly acknowledge the limits in how well a tool can fulfill its function; yet investigators who study change wrestle with systematic questions about the role of inference when a tool is used to describe dynamic events. Differences alone are not usually accepted as a sign of development, and strong measurement approaches allow researchers occasionally to find surprises in how their instruments record events. Ideally speaking, measurement surprises can potentially enhance rather than undermine valid descriptions of predictable forms of change.

A third set of measurement challenges emerged once researchers disagreed with the emphasis placed on studying only concepts that can be predictably generalized across time, contexts, or persons. These interpretive theorists focus on events that might be fleeting and unpredictable. Interpretive measurement paradigms are used to capture immediate contexts and reactions to them accurately and consistently with little or no concern about whether these events can be replicated. Using a wide range of methods, investigators relying on interpretive frameworks offer rich descriptions of highly complex contexts as they endeavor to draw accurate inferences about their observations. Like psychometricians and developmentalists, researchers who adopt interpretive measurement methods rely on rule-governed logic. Nevertheless, interpretive theorists focus on how well their evidence corresponds to systematic philosophical rules. Strong validity arguments are generated using a philosophical base that highlights the degree to which data are collected in a dependable and credible manner. Validity is established if others can come to the same conclusions about the available data that were generated by the agent conducting research.

Interpretive theorists differ widely in the types of inferences they defend. Case studies, ethnographies, and ethnologies, for example, use information from multiple sources in support of inferences, but they differ in how many levels of inferences are supported. Various levels of inference are defended by drawing on concepts inherent in philosophical traditions such as phenomenology, critical theory, interpretive analytics, and deconstructionism. Validity arguments are made and evaluated by deferring to the philosophical rules that govern these traditions.

Put simply, investigators working with different measurement approaches focus on very different research problems. Psychometricians endeavor to describe accurately tangible constructs. Developmentalists seek to understand predictable forms of change, and interpretive theorists endeavor to understand how everyday events align with the tenets of various philosophical traditions. These agendas play a profound role in how validity arguments are formed and in how researchers identify threats to such validity.

## Establishing Validity

Because investigators are interested in different phenomena, language for representing the dimensions of psychometric, developmental, and interpretive validity arguments might vary. Nevertheless, all measurement arguments hinge on a few basic principles. Investigators show evidence of how well indicants of the phenomenon under consideration actually represent the ideas to be measured and whether the indicants are recorded with precision. They also evaluate the comprehensiveness and balance of the measurement instrument, and whether the collection of evidence generated actually supports the descriptions offered, and only those descriptions.

### *Psychometric Concerns*

In the psychometric tradition, the language for representing validity is clear but abstract. Using probability theory as a guide, investigators look at the content of the instrument to determine whether the tool adequately samples the whole of the construct to be measured. *Content analysis* involves procedures for analyzing each indicant, establishing reliability, and evaluating the representativeness of selected items in relation to the whole of what is to be measured.

*Criterion analysis* extends the process of evaluating the content of a measure to document how well the indicants on a tool adequately represent each dimension of a construct and only that dimension. Dimensions might be related to one another, and this is established with arguments about *concurrent validity.* Dimensions might also differ from one another while remaining components of a complex construct. *Discriminant validity* arguments are used to establish the importance of such potentially contrasting perspectives to the measurement of a construct.

Finally, *construct analysis* involves comparing the construct to be measured with other existing constructs to verify its relative uniqueness. *Convergent validity* is established by comparing a new instrument with other related instruments to verify that they measure the same construct. *Divergent validity* is established by comparing a new instrument with unrelated instruments to verify that these in fact are distinct. Strong validity arguments consider all of these challenges while acknowledging that alterations at one level affect the validity of all other levels, leading to imperfections in how even well-designed tools function.

### *Developmental Concerns*

Language for measuring change can seem quite different, and this difference reflects the concern with how well an instrument assesses the process of change. Parallel to the psychometric concern with content validity, strong developmental arguments include information on a tool's *calibration.* When investigators want to measure structural change, they look at how well the instrument functions with and without externally imposed constraints. Calibration for measuring temporal change is adjusted according to whether investigators want to measure directly the rate, magnitude, or duration of change as well as address stability, order, and randomness. Validity arguments also distinguish structural and temporal types of change from the measurement of individual differences or diversity across persons.

Much like psychometricians' concerns about criterion validity, developmentalists also evaluate the degree of *standardization* required across measurement events. They recognize that some forms of standardization are trivial and can undermine the measurement of change, but other forms of consistency are essential for documenting the change under investigation. Justifying these decisions and documenting key assumptions about time, maturity, and the integration of various change dimensions are crucial for establishing that an instrument does an adequate job of representing the processes under investigation and only those processes.

The third validity issue, *generalizability,* is the most unlike its psychometric counterpart of construct validity in that developmentalists are defending the assessment of a process and not a tangible construct. It can be confusing to imagine how change is replicated well enough to support broader conclusions about its dynamic nature, but such conclusions are possible if an instrument is calibrated well enough to offer a consistent representation of intended transformations. Developmentalists might draw general conclusions about the order with which some events occur, the degree of synchrony with which change occurs, or how well the dimensions under investigation follow structural patterns. They also draw conclusions about whether the findings obtained with a measurement tool are intended to represent normative phenomena that would be universal for all humankind or contingent phenomena that would be apparent only for some settings or some persons.

### *Interpretative Concerns*

The same basic principles that are evident for other forms of measurement are altered dramatically but present in interpretive approaches as well. Investigators seek to *minimize bias* in how evidence is gathered. They do not compare indicators of the measured phenomena to definitions of

constructs or of change. Instead, they illustrate how evidence is selected accurately to represent the full range of possible perspectives and contextual features while constraining investigators' inevitably biased judgments. Rules that govern different philosophical traditions highlight various forms of bias and help investigators see beyond their own social position well enough to draw well-justified conclusions about the evidence before them. Investigators also establish *dependability* and *credibility* in how the gathered evidence is aligned with the interpretations they make. In doing so, they systematically evaluate themselves, other informants, and any artifacts that are gathered to support the inferences in question. Finally, each inference is evaluated for its relative *confirmability.* That is, investigators evaluate whether each form of evidence meets criteria established for *descriptive validity, interpretive validity, theoretical validity, generalizability,* and *evaluative validity.* This confirmation involves a process of systematically labeling key ideas, comparing descriptions with actual events, critiquing the performance of the researchers, evaluating the quality of any interpersonal relationships, and evaluating the overall persuasiveness of the evidence used to support the existence of the inferences being measured.

### *A Validity Argument*

Although this overview offers only a superficial account of how investigators working within different measurement traditions support their case for validity, it should be easy to see that establishing validity is not a simple process. It is helpful to remember that validity in measurement is an inherently creative act that often involves a process of trial and error.

## Limiting Threats to Validity

Despite the creativity involved in producing a strong validity argument, common threats to validity can be anticipated. At each step in the measurement process, there are opportunities for instruments and those who use them to fall short of the ideal. Although for each measurement tradition, there are a myriad of rules and procedures for maximizing validity, most can be encompassed by three general guidelines.

### *Define Clear Measurement Aims*

This first guideline for establishing measurement goals might seem obvious, but it is easy for investigators to deviate, sometimes quite radically, from their initial agendas as they design a new measurement tool. Measurement aims typically include at least two basic concerns, and validity is compromised if one or both are overlooked. First, regardless of their preferred measurement approach, investigators articulate the theoretical framework that will guide their inquiry. Second, investigators define a level of generality they hope to achieve with their research. Some investigators focus on narrowly defined or micro-level concepts or inference, whereas other investigators adopt a macro-level lens through which to define what they hope to assess. Investigators also hold different assumptions about whether their target concepts are universal, normative, or contingent on situational factors. Together, decisions about the theoretical framework and levels of generality affect recommendations for who and what should be sampled, and sample characteristics, in turn, play a profound role in the generation of strong validity arguments.

### *Compare Obtained Evidence With an Ideal*

Again, each measurement approach works with a different understanding of the ideal, but in each case, validity is threatened when investigators forget to compare the evidence they have obtained or might obtain in the measurement process with an understanding of what they would like to assess. Using several well-designed methods, investigators endeavor to offer a balanced representation of the concepts they are measuring and minimize distortion in how the concepts are represented. Psychometricians, for example, commonly use theoretical blueprints and compare the indicants of a tool with the categories on the blueprint. Developmentalists often rely on a process of introducing constraints and removing them to predict and control the change they hope to understand. Interpretive theorists evaluate the characteristics of their informants, the artifacts, and a host of contextual factors to determine whether they have gathered all the relevant evidence, making decisions about what evidence to retain in light of philosophical rules. In short, everyone looks for distortion in